Practical methods for modelling weak VARMA processes: identification, estimation and specification with a macroeconomic application*

Jean-Marie Dufour[†]
McGill University

Denis Pelletier[‡]
North Carolina State University

October 2008

*The authors thank Marine Carasco, John Galbraith, Nour Meddahi and Rui Castro for several useful comments. The second author gratefully acknowledges financial assistance from the Social Sciences and Humanities Research Council of Canada, the Government of Québec (fonds FCAR), the CRDE and CIRANO. Earlier versions of the paper circulated under the title *Linear Estimation of Weak VARMA Models With a Macroeconomic Application*. This work was supported by the Social Sciences and Humanities Research Council of Canada, the Natural Sciences and Engineering Research Council of Canada, the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Canada Council for the Arts (Killam Fellowship), the CIREQ, the CIRANO, and the Fonds FCAR (Government of Québec).

[†] William Dow Professor of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ). Mailing address: Department of Economics, McGill University, Leacock Building, Room 519, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada. TEL: (1) 514 398 8879; FAX: (1) 514 398 4938; e-mail: jean-marie.dufour@mcgill.ca . Web page: http://www.jeanmariedufour.com

[‡] Department of Economics, Box 8110, North Carolina State University, Raleigh, NC 27695-8110, USA. Email: denis_pelletier@ncsu.edu. Web page: http://www4.ncsu.edu/ dpellet

ABSTRACT

In this paper, we develop practical methods for modelling weak VARMA processes. In a first part, we propose new identified VARMA representations, the *diagonal MA equation form* and the *final MA equation form*, where the MA operator is diagonal and scalar respectively. Both of these representations have the important feature that they constitute relatively simple modifications of a VAR model (in contrast with the echelon representation). In a second part, we study the problem of estimating VARMA models by relatively simple methods which only require linear regressions. We consider a generalization of the regression-based estimation method proposed by Hannan and Rissanen (1982). The asymptotic properties of the estimator are derived under weak hypotheses on the innovations (uncorrelated and strong mixing) so as to broaden the class of models to which it can be applied. In a third part, we present a modified information criterion which gives consistent estimates of the orders under the proposed representations. To demonstrate the importance of using VARMA models to study multivariate time series we compare the impulse-response functions and the out-of-sample forecasts generated by VARMA and VAR models.

Key words: linear regression; VARMA; final equation form; information criterion; weak representation; strong mixing condition; impulse-response function.

Journal of Economic Literature Classification: C13, C32, C51, E0.

1. Introduction

In time series analysis and econometrics, VARMA models are scarcely used to represent multivariate time series. VAR models are much more widely employed because they are easier to implement. The latter models can be estimated by least squares methods, while VARMA models typically require nonlinear methods (such as maximum likelihood). Specification is also easier for VAR models since only one lag order must be chosen.

VAR models, however, have important drawbacks. First, they are typically less parsimonious than VARMA models [*e.g.*, see Lütkepohl and Poskitt (1996b)]. Second, the family of VAR models is not closed under marginalization and temporal aggregation [see Lütkepohl (1991)]. The truth cannot always be a VAR. If a vector satisfies a VAR model, subvectors do not typically satisfy VAR models (but VARMA models). Similarly, if the variables of a VAR process are observed at a different frequency, the resulting process is not a VAR process. In contrast, the class of VARMA models is closed under such operations.

The importance of nonlinear models has been growing in the time series literature. These models are interesting and useful but may be hard to use. Because of this and the fact that many important classes of nonlinear processes admit an ARMA representation [*e.g.*, see Francq and Zakoïan (1998), Francq, Roy, and Zakoïan (2003)] many researchers and practitioners still have an interest in linear ARMA models. However, the innovations in these ARMA representations do not have the usual i.i.d. or m.d.s. property, although they are uncorrelated. One must then be careful before applying usual results to the estimation of ARMA models because they usually rely on the above strong assumptions [*e.g.*, see Brockwell and Davis (1991) and Lütkepohl (1991)]. We refer to these as strong and semi-strong ARMA models respectively, by opposition to weak ARMA models where the innovations are only uncorrelated. The i.i.d. and m.d.s. properties are also not robust to aggregation (the i.i.d. Gaussian case being an exception); see Francq and Zakoïan (1998), Francq, Roy, and Zakoïan (2003), Palm and Nijman (1984), Nijman and Palm (1990), Drost (1993). In fact, the Wold decomposition only guarantees that the innovations are uncorrelated.

It follows that (weak) VARMA models appear to be preferable from a theoretical viewpoint, but their adoption is complicated by identification and estimation difficulties. The direct multivariate generalization of ARMA models does not give an identified representation [see Lütkepohl (1991, Section 7.1.1)]. It follows that one has to decide on a set of constraints to impose so as to achieve identification. Standard estimation methods for VARMA models (maximum likelihood, nonlinear least squares) require nonlinear optimization which may not be feasible as soon as the model involves a few time series, because the number of parameters can increase quickly.

In this paper, we consider the problem of modeling weak VARMA processes. Our goal is to develop a procedure which will ease the use of these models. It will cover three basic modelling operations: identification, estimation and specification.

First, in order to avoid identification problems and to further ease the use of VARMA models, we introduce three new identified VARMA representations, the *diagonal MA equation form*, the *final MA equation form* and the *diagonal AR equation form*. Under the diagonal MA equation form (diagonal AR equation form) representation, the MA (AR) operator is diagonal and each lag operator may have a different order. Under the final MA equation form representation the MA operator is scalar, *i.e.* the operators are equal across equations. The diagonal and final MA equation form representations can be interpreted as simple extensions of the VAR model, which should be appealing to practitioners who prefer to employ VAR models due to their ease of use. The identified VARMA representation which is the most widely employed in the literature is the *echelon form*. Specification of VARMA models in echelon form does not amount to specifying the order p and q as with ARMA models. Under this representation, VARMA models are specified by as many parameters, called Kronecker indices, as the number of time series studied. These indices determine the order of the elements of the AR and MA operators in a non trivial way. The complicated nature of the echelon form representation is a major reason why practitioners are not using VARMA models, so the introduction of a simpler identified representation is interesting.

Second, we consider the problem of estimating VARMA models by relatively simple methods which only require linear regressions. For that purpose, we consider a multivariate generalization of the regression-based estimation method proposed by Hannan and Rissanen (1982) for univariate ARMA models. The method is performed in three steps. In a first step, a long autoregression is fitted to the data. In the second step, the lagged innovations in the ARMA model are replaced by the corresponding residuals from the long autoregression and a regression is performed. In a third step, the data from the second step are filtered so as to give estimates that have the same asymptotic covariance matrix than one would get with the maximum likelihood [claimed in Hannan and Rissanen (1982), proven in Zhao-Guo (1985)]. Extension of this innovation-substitution method to VARMA models was also proposed by Hannan and Kavalieris (1984a) and Koreisha and Pukkila (1989), under the assumption that the innovations are a m.d.s.

Here, we extend these results by showing that the linear regression-based estimators are consistent under weak hypotheses on the innovations and how filtering in the third step gives estimators that have the same asymptotic distribution as their nonlinear counterparts (maximum likelihood if the innovations are i.i.d., or nonlinear least squares if they are merely uncorrelated). In the non i.i.d. case, we consider strong mixing conditions [Doukhan (1995), Bosq (1998)], rather than the usual m.d.s. assumption. By using weaker assumptions for the process of the innovations, we broaden the class of processes to which our method can be applied.

Third, we suggest a modified information criterion to choose the orders of VARMA models under these representations. This criterion is to be minimized in the second step of the estimation method over the orders of the AR and MA operators and gives consistent estimates of these orders. Our criterion is a generalization of the information criterion proposed by Hannan and Rissanen (1982), which was later corrected by Hannan and Rissanen (1983) and Hannan and Kavalieris (1984b), for choosing the orders p and q in ARMA models. The idea of generalizing this information criterion is mentioned in Koreisha and Pukkila (1989) but a specific generalization and theoretical properties are not presented.

Fourth, the method is applied to U.S. macroeconomic data previously studied by Bernanke and Mihov (1998) and McMillin (2001). To illustrate the impact of using VARMA models instead of VAR models to study multivariate time series we compare the impulse-response functions generated by each model. We show that we can obtain much more precise estimates of the impulse-response function by using VARMA models instead of VAR models.

The rest of the paper is organized as follows. Our framework and notation are described in section 2. The new identified representations are presented in section 3. In section 4, we present the estimation method. In section 5, we describe the information criterion used for choosing the orders of VARMA models under the representation proposed in our work. Section 6 contains results of Monte Carlo simulations which illustrate the properties of our method. Section 7 presents the macroeconomic application where we compare the impulse-response functions from a VAR model and VARMA models. Section 8 contains a few concluding remarks. Finally, proofs are in the appendix.

2. Framework

Consider the following K-variate zero mean VARMA(p,q) model in standard representation:

$$Y_t = \sum_{i=1}^{p} \Phi_i Y_{t-i} + U_t - \sum_{j=1}^{q} \Theta_j U_{t-j} \,,\ t \in \mathbb{Z}, \tag{2.1}$$

where U_t is a sequence of uncorrelated random variables with mean zero, defined on some probability space $(\Omega,\, \mathcal{A},\, \mathcal{P})$. The vectors Y_t and U_t contain the K univariate time series: $Y_t = [y_{1t}, y_{2t},\, \ldots\, ,\, y_{Kt}]'$ and $U_t = [u_{1t}, u_{2t},\, \ldots\, ,\, u_{Kt}]'$. We can also write the previous equation with lag operators:

$$\Phi(L)Y_t = \Theta(L)U_t \tag{2.2}$$

where

$$\Phi(L) = I_K - \Phi_1 L - \cdots - \Phi_p L^p \,, \quad \Theta(L) = I_K - \Theta_1 L - \cdots - \Theta_q L^q . \tag{2.3}$$

Let $\mathcal{H}_t$ be the Hilbert space generated by $(Y_s,\ s < t)$. The process U_t can be interpreted as the linear innovation of Y_t:

$$U_t = Y_t - \mathsf{E}_L[Y_t|\mathcal{H}_t]. \tag{2.4}$$

We assume that Y_t is a strictly stationary and ergodic sequence and that the process U_t has common variance ($Var[U_t] = \Sigma_U$) and finite fourth moment ($E[|u_{it}|^{4+2\delta}] < \infty$, for all i and t, where $\delta > 0$). We make the zero mean-mean hypothesis only to simplify notation.

Assuming that the process Y_t is stable,

$$\det\left[\Phi(z)\right] \neq 0 \text{ for all } |z| \leq 1 \,, \tag{2.5}$$

and invertible,

$$\det\left[\Theta(z)\right] \neq 0 \text{ for all } |z| \leq 1 \,, \tag{2.6}$$

it can be represented as an infinite VAR,

$$\Pi(L)Y_t = U_t \tag{2.7}$$

where

$$\Pi(L) = \Theta(L)^{-1}\Phi(L) = I_K - \sum_{i=1}^{\infty} \Pi_i L^i, \tag{2.8}$$

or an infinite VMA,

$$Y_t = \Psi(L) U_t \tag{2.9}$$

where

$$\Psi(L) = \Phi(L)^{-1}\Theta(L) = I_K - \sum_{j=1}^{\infty} \Psi_j L^j. \tag{2.10}$$

We will denote by $\varphi_{ij}(L)$ the polynomial in row i and column j of $\Phi(L)$, and the row i or column j of $\Phi(L)$ by

$$\Phi_{i\bullet}(L) = [\varphi_{i1}(L), \ldots, \varphi_{iK}(L)], \tag{2.11}$$

$$\Phi_{\bullet j}(L) = [\varphi_{1j}(L), \ldots, \varphi_{Kj}(L)]'. \tag{2.12}$$

The diag operator creates a diagonal matrix,

$$\mathrm{diag}[\varphi_{11}(L), \ldots, \varphi_{KK}(L)] = \begin{bmatrix} \varphi_{11}(L) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi_{KK}(L) \end{bmatrix} \tag{2.13}$$

where

$$\varphi_{ii}(L) = 1 - \varphi_{ii,1}L - \cdots - \varphi_{ii,p}L^p. \tag{2.14}$$

The function $\deg[\varphi(L)]$ returns the degree of the polynomial $\varphi(L)$ and the function $dim(\gamma)$ gives the dimension of the vector γ.

We need to impose some structure on the process U_t. The typical hypothesis which is imposed in the time series literature is that the U_t's are either independent and identically distributed (i.i.d.) or a martingale difference sequence (m.d.s.). In this work, we do not impose such strong assumptions because we want to broaden the class of models to which it can be applied. We only assume that it satisfies a strong mixing condition [Doukhan (1995), Bosq (1998)]. Let U_t be a strictly stationary process, and

$$\alpha(h) = \sup_{\substack{B \in \sigma(U_s, s \le t) \\ C \in \sigma(U_s, s \ge t+h)}} |\Pr(B \cap C) - \Pr(B)\Pr(C)| \tag{2.15}$$

the α-mixing coefficient of order $h \geq 1$, where $\sigma(U_s, s \le t)$ and $\sigma(U_s, s \ge t+h)$ are the σ-algebras associated with $\{U_s : s \le t\}$and $\sigma(U_s : s \ge t + h)$ respectively. We suppose that U_t is strong mixing, *i.e.*

$$\sum_{h=1}^{\infty} \alpha(h)^{\delta/(2+\delta)} < \infty \quad \text{for some} \quad \delta > 0. \tag{2.16}$$

This is a fairly minimal condition that will be satisfied by many processes of interest.

3. Identification and diagonal VARMA representations

It is important to note that we cannot work with the standard representation (2.1) because it is not identified. To help gain intuition on the identification of VARMA models, we can consider a more general representation where Φ_0 and Θ_0 are not identity matrices:

$$\Phi_0 Y_t = \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + \Theta_0 U_t - \Theta_1 U_{t-1} + \cdots + \Theta_q U_{t-q}. \tag{3.1}$$

By this specification, we mean the well-defined process

$$Y_t = (\Phi_0 - \Phi_1 L - \cdots - \Phi_p L^p)^{-1} (\Theta_0 + \Theta_1 L + \cdots + \Theta_q L^q)\, U_t. \tag{3.2}$$

But we can see this such process has a standard representation if Φ_0 and Θ_0 are nonsingular. To see this, we premultiply (3.1) by Φ_0^{-1} and define $\bar{U}_t = \Phi_0^{-1} \Theta_0 U_t$:

$$\begin{aligned} Y_t \;=\;& \Phi_0^{-1} \Phi_1 Y_{t-1} + \cdots + \Phi_0^{-1} \Phi_p Y_{t-p} \\ & + \bar{U}_t - \Phi_0^{-1} \Theta_1 \Theta_0^{-1} \Phi_0 \bar{U}_{t-1} - \cdots - \Phi_0^{-1} \Theta_q \Theta_0^{-1} \Phi_0 \bar{U}_{t-q}. \end{aligned} \tag{3.3}$$

Redefining the matrices, we get a representation of type (2.1). As long as Φ_0 and Θ_0 are nonsingular, we can transform a non-standard VARMA into a standard one.

We say that two VARMA representations are equivalent if $\Phi(L)^{-1}\Theta(L)$ results in the same operator $\Psi(L)$. Thus, to ensure uniqueness of a VARMA representation, we must impose restrictions on the AR and MA operators such that for a given $\Psi(L)$ there is one and only one set of operators $\Phi(L)$ and $\Theta(L)$ that can generate this infinite MA representation.

A first restriction that we impose is a multivariate equivalent of the coprime property in the univariate case. We do not want factors of $\Phi(L)$ and $\Theta(L)$ to "cancel out" when $\Phi(L)^{-1}\Theta(L)$ is computed. This feature is called the *left-coprime* property [see Hannan (1969) and Lütkepohl (1993)]: the matrix operator $\Psi[\Phi(L), \Theta(L)] \equiv \Phi(L)^{-1}\Theta(L)$ is *left-coprime* if, for any operators $D(L)$, $\bar{\Phi}(L)$, and $\bar{\Theta}(L)$, the identity

$$D(L)\Psi[\bar{\Phi}(L), \bar{\Theta}(L)] = \Psi[\Phi(L), \Theta(L)] \tag{3.4}$$

implies that $D(L)$ is *unimodular* [*i.e.*, $\det D(L)$ is a nonzero constant]. To obtain uniqueness of left-coprime operators we have to impose restrictions ensuring that the only feasible unimodular operator $D(L)$ in (3.4) is $D(L) = I_K$. There is no unique way of doing this. The dominant representation in the literature is the *echelon form* [see Deistler and Hannan (1981), Hannan and Kavalieris (1984b), Lütkepohl (1993), Lütkepohl and Poskitt (1996a)].

Definition 3.1 (Echelon form) *The VARMA representation in* (2.1) *is said to be in echelon form if the AR and MA operators* $\Phi(L) = [\varphi_{ij}(L)]_{i,j=1,\ldots,K}$ *and* $\Theta(L) = [\theta_{ij}(L)]_{i,j=1,\ldots,K}$ *satisfy the following conditions: all operators* $\varphi_{ij}(L)$ *and* $\theta_{ij}(L)$ *in the i-th row of* $\Phi(L)$ *and* $\Theta(L)$ *have the* *same degree p_i with the form*

$$\begin{aligned}
\varphi_{ii}(L) &= 1 - \sum_{m=1}^{p_i} \varphi_{ii,m} L^m, \quad \textit{for } i = 1, \dots, K, \\
\varphi_{ij}(L) &= - \sum_{m=p_i-p_{ij}+1}^{p_i} \varphi_{ij,m} L^m, \quad \textit{for } j \neq i, \\
\theta_{ij}(L) &= \sum_{m=0}^{p_i} \theta_{ij,m} L^m \quad \textit{for } i, j = 1, \dots, K, \quad \textit{with } \Theta_0 = \Phi_0.
\end{aligned}$$

Further, in the VAR operator $\varphi_{ij}(L)$,

$$p_{ij} = \begin{cases} \min(p_i + 1, p_j) & \textit{for } i \geq j, \\ \min(p_i, p_j) & \textit{for } i < j, \end{cases}$$

$i, j = 1, \dots, K$, i.e. p_{ij} *specifies the number of free coefficients in the operator $\varphi_{ij}(L)$ for $j \neq i$. The row orders $(p_1, \dots, p_K)$ are the Kronecker indices and their sum $\sum_{i=1}^{K} p_i$ is the McMillan degree. For the VARMA orders we have in general $p = q = \max(p_1, \dots, p_K)$.*

We see that dealing with VARMA models in echelon form is not as easy as dealing with univariate ARMA models where everything is specified by choosing the value of p and q. The number of Kronecker indices is larger than two (if K is larger than two) and, when choosing p_{ij}, we have to consider if we are above or below the diagonal. Having a summation subscript in the operator φ_{ij}, $m = p_i - p_{ij} + 1$, different across rows and columns also complicates the use of this representation. The task is far from being impossible but it is more complicated than for ARMA models. Specification of VARMA models in echelon form is discussed in Hannan and Kavalieris (1984b), Lütkepohl and Claessen (1997), Poskitt (1992), Nsiri and Roy (1992, 1996), Lütkepohl and Poskitt (1996b), Bartel and Lütkepohl (1998). This might be a reason why practitioners are reluctant to employ VARMA models. Who could blame them for sticking with VAR models when they probably need to refer to a textbook to simply write down an identified VARMA representation?

In this work, to ease the use of VARMA models we present new VARMA representations which can be seen as a simple extensions of the VAR model. To introduce them, we first review another identified representation, the *final equation form*, which will refer to as the *final AR equation form*, under which the AR operator is scalar [see Zellner and Palm (1974), Hannan (1976), Wallis (1977), Lütkepohl (1993)].

Definition 3.2 (Final AR equation form) *The VARMA representation* (2.1) *is said to be in final AR equation form if $\Phi(L) = \varphi(L) I_K$, where $\varphi(L) = 1 - \varphi_1 L - \cdots - \varphi_p L^p$ is a scalar polynomial with $\varphi_p \neq 0$.*

To see how we can obtain a VARMA model with a final AR equation form representation, we can proceed as follows. By standard linear algebra, we have

$$\Phi(L)^{*}\Phi(L) = \Phi(L)\Phi(L)^{*} = \det\left[\Phi(L)\right] I_K \tag{3.5}$$

where $\Phi(L)^{*}$ is the adjoint matrix of $\Phi(L)$. On multiplying both sides of (2.2) by $\Phi(L)^{*}$, we get:

$$\det\left[\Phi(L)\right] Y_t = \Phi(L)^{\star}\Theta(L)U_t. \tag{3.6}$$

This representation is not attractive for several reasons. First, it is quite far from usual VAR models by excluding lagged values of other variables in each equation (*e.g.*, the AR part of the first equation includes lagged values of y_{1t} but no lagged values of $y_{2t}, \ldots, y_{Kt}$). Further, the AR coefficients are the same in all the equations, which will require a polynomial of higher order pK. Second, the interaction between the different variables is modeled through the MA part of the model, which may have to be quite complex.

However, we can derive alternative representations which are both more intuitive and practical. First, upon multiplying both sides of (2.2) by $\Theta(L)^{\star}$, we get:

$$\Theta(L)^{\star}\Phi(L)Y_t = \det\left[\Theta(L)\right] U_t \tag{3.7}$$

where $\Theta(L)^{\star}$ is the adjoint matrix of $\Theta(L)$. We refer to VARMA models in (3.7) as being in *final MA equation form*.

Definition 3.3 (Final MA equation form) *The VARMA representation* (2.1) *is said to be in final MA equation form if*

$$\Theta(L) = \theta(L)I_K \tag{3.8}$$

where $\theta(L) = 1 - \theta_1 L - \cdots - \theta_q L^q$ *is a scalar operator with* $\theta_q \neq 0$.

By (3.7), it is clear that any VARMA process satisfying (2.1) - (2.6) can be written in final MA form. This form is much closer to the usual finite-order VAR model than the echelon representation or the final AR equation form, because the AR part is a finite-order VAR while the MA part of each equation is a *univariate* MA which only involves a single innovation process. The main drawback comes from the fact that the MA operator is the same in all the equations, which can lead to a high-order MA. It is however possible to get a more parsimonious representation by allowing for different MA polynomials in different equations.

Suppose there are common roots across rows for some columns of $\Theta(L)$, so that starting from (2.1) we can write:

$$\Phi(L)Y_t = \bar{\Theta}(L)D(L)U_t, \tag{3.9}$$

$$\bar{\Theta}(L)^{\star}\Phi(L)Y_t = \det\left[\bar{\Theta}(L)\right] D(L)U_t, \tag{3.10}$$

where $D(L) = \mathrm{diag}[d_1(L), \ldots, d_K(L)]$ and $d_j(L)$ is a polynomial common to $\theta_{ij}(L)$, $\forall i = 1, \ldots, K$. We see that allowing for diagonal polynomials in the moving average as in equation (3.10) may yield a more parsimonious representation than (3.7). We will call the representation (3.10) *diagonal MA equation form* representation.

Definition 3.4 (Diagonal MA equation form) *The VARMA representation* (2.1) *is said to be in diagonal MA equation form if* $\Theta(L) = \mathrm{diag}[\theta_{ii}(L)] = I_K - \Theta_1 L - \cdots - \Theta_q L^q$ *where* $\theta_{ii}(L) = 1 - \theta_{ii,1}L - \cdots - \theta_{ii,q_i}L^{q_i}$, $\theta_{ii,q_i} \neq 0$ *and* $q = \max_{1\leq i\leq K}(q_i)$.

The latter representation is interesting because contrary to the echelon form it is relatively easy to specify. We do not have to deal with rules for the orders of the off-diagonal elements in the AR and MA operators. The fact that it can be seen as a simple extension of the VAR model is also appealing. Practitioners are comfortable using VAR models, so simply adding lags of u_{it} to equation i is a natural extension of the VAR model which could give a more parsimonious representation. It also has the advantage of putting the simple structure on the MA polynomials, the part which complicates the estimation, rather than the AR part as in the final AR equation form. Notice that in VARMA models, it is not necessary to include lags of all the innovations $u_{1t}, \cdots, u_{Kt}$ in every equation. This could entice practitioners to consider VARMA models if it is combined with a simple regression-based estimation method.

From (3.7). it is clear that any process that satisfies (2.1) - (2.6) also possesses a diagonal MA representation (because the latter includes the final MA equation form as a special case). We will now give conditions ensuring that a diagonal MA representation is unique. For that purpose, we consider the following assumptions and use the following matrix lemma (which may be of separate interest).

Assumption 3.5 *The matrices* $\Phi(z)$ *and* $\Theta(z)$ *have the following form:*

$$\Phi(z) = I_K - \Phi_1 z - \cdots - \Phi_p z^p\,, \quad \Theta(z) = I_K - \Theta_1 z - \cdots - \Theta_q z^q.$$

Assumption 3.6 $\Theta(z)$ *is diagonal:*

$$\Theta(z) = \mathrm{diag}\left[\theta_{11}(z),\, \ldots\,,\, \theta_{KK}(z)\right]$$

where $\theta_{ii}(z) = 1 - \theta_{ii,1}z - \cdots - \theta_{ii,q_i}z^{q_i}$ *and* $\theta_{ii,q_i} \neq 0$, $i = 1, \ldots, K$.

Assumption 3.7 *For each* $i = 1, \ldots, K$, *there are no roots common to* $\Phi_{i\bullet}(z)$ *and* $\theta_{ii}(z)$, i.e. *there is no value* $z^\star$ *such that* $\Phi_{i\bullet}(z^\star) = 0$ *and* $\theta_{ii}(z^\star) = 0$.

Lemma 3.8 *Let* $[\Phi(z), \Theta(z)]$ *and* $\left[\bar{\Phi}(z), \bar{\Theta}(z)\right]$ *be two pairs of polynomial matrices which satisfy the Assumptions* **3.5** *to* **3.7**. *If*

$$\Phi(z)^{-1}\Theta(z) = \bar{\Phi}(z)^{-1}\bar{\Theta}(z)\,,\ \text{for } 0 \leq |z| < \rho_0, \tag{3.11}$$

where ρ_0 *is a positive constant, then*

$$\Phi(z) = \bar{\Phi}(z) \text{ and } \Theta(z) = \bar{\Theta}(z)\,,\ \forall z. \tag{3.12}$$

The proof of this lemma as well as other propositions appear in the Appendix. It entails that the matrix $\Phi(z)^{-1}\Theta(z)$ has a unique factorization in terms of polynomial matrices $\Phi(z)$ and $\Theta(z)$, *i.e.*

the operators $\Phi(L)$ and $\Theta(L)$ are uniquely defined by $\Phi(L)^{-1}\Theta(L)$. It is also easy to see that the condition

$$\Phi(z)^{-1}\Theta(z) = \bar{\Phi}(z)^{-1}\bar{\Theta}(z) \tag{3.13}$$

could be replaced by

$$\Theta(z)^{-1}\Phi(z) = \bar{\Theta}(z)^{-1}\bar{\Phi}(z) \tag{3.14}$$

since by assumption the inverses of $\Theta(z)$ and $\bar{\Theta}(z)$ exist. Note that Assumption **3.5** is equivalent to (2.3). It is interesting to note that the conditions of Lemma **3.8** allow $\det[\Phi(z)]$ and $\det[\Theta(z)]$ to have roots on or inside the unit circle $|z| = 1$. Further, Assumption **3.7** is weaker than the hypothesis that $\det[\Phi(L)]$ and $\det[\Theta(L)]$ have no common roots, which would be a generalization of the usual identification condition for ARMA models. We can now show that a VARMA model in diagonal MA form has a unique representation.

Theorem 3.9 (Identification of diagonal MA equation form representation) *Let* $\{Y_t : t \in \mathbb{Z}\}$ *be a VARMA process satisfying the conditions* (2.1) - (2.6). *If the assumptions* **3.6** *and* **3.7** *hold, then the polynomial operators* $\Phi(L)$ *and* $\Theta(L)$ *are uniquely defined.*

Similarly, we can demonstrate that the final MA equation form representation is identified under the following assumption.

Assumption 3.10 *There are no roots common to* $\Phi(z)$ *and* $\theta(z)$, i.e. *there is no value* $z^{\star}$ *such that* $\Phi(z^{\star}) = 0$ *and* $\theta(z^{\star}) = 0$.

Theorem 3.11 (Identification of final MA equation form representation) *Let* $\{Y_t : t \in \mathbb{Z}\}$ *be a VARMA process satisfying the conditions* (2.1) - (2.6). *If the model is in final MA equation form and Assumption* **3.7** *holds, then the polynomial operators* $\Phi(L)$ *and* $\Theta(L)$ *are uniquely defined.*

From equation (3.7), we see that it is always possible to obtain a diagonal MA equation form representation starting from any VARMA representation. One case where we would obtain a diagonal and not final MA representation is when there are common factors across rows of columns of $\Theta(L)$ as in (3.10).

A strong appeal of the diagonal and final MA equation form representations is that it is easy to get the equivalent (in term of autocovariances) invertible MA representation of a non-invertible representation. With ARMA models, we simply have to invert the roots of the MA polynomial which are inside the unit circle and adjust the standard deviation of the innovations (divide it by the square of these roots): see Hamilton (1994, Section 3.7). The same procedure could be applied to VARMA models in diagonal or final MA equation form.

For VARMA representations where no particular simple structure is imposed on the MA part, at the moment we are not aware of an algorithm to go from the non-invertible to the invertible representation tough theoretically this invertible representation exist and is unique as long as $\det[\Theta(z)] \neq 0$ for $|z| = 1$; see Hannan and Deistler (1988, chapter 1, section 3). So it might be troublesome to use a nonlinear optimization with these VARMA representations since we don't know how to go from the non-invertible to the invertible representation.

We can also consider the following natural generalization of the final AR equation form, where we simply replace the scalar AR operator by a diagonal operator.

Definition 3.12 (Diagonal AR equation form) *The VARMA model*(2.1) *is said to be in diagonal AR equation form if* $\Phi(L) = \mathrm{diag}[\varphi_{ii}(L)] = I_K - \Phi_1 L - \cdots - \Phi_p L^p$ *where* $\varphi_{ii}(L) = 1 - \varphi_{ii,1}L - \cdots - \varphi_{ii,p_i}L^{p_i}$ *and* $p = \max_{1\leq i\leq K}(p_i)$.

Assumption 3.13 *For each* $i = 1, \ldots, K$, *there are no roots common to* $\varphi_{ii}(z)$ *and* $\Theta_{i\bullet}(z)$, i.e. *there is no value* $z^{\star}$ *such that* $\varphi_{ii}(z^{\star}) = 0$ *and* $\Theta_{i\bullet}(z^{\star}) = 0$.

Theorem 3.14 (Identification of diagonal AR equation form representation) *Let* $\{Y_t : t \in \mathbb{Z}\}$ *be a VARMA process satisfying the conditions* (2.1) - (2.6). *If the model is in diagonal AR equation form and Assumption* **3.13** *holds, then the polynomial operators* $\Phi(L)$ *and* $\Theta(L)$ *are uniquely defined.*

From Theorem **3.9**, we can see that one way to ensure identification is to impose constraints on the MA operator. This is an alternative approach to the ones developed for example in Hannan (1971, 1976) where the identification is obtained by restricting the autoregressive part to be lower triangular with $deg[\varphi_{ij}(L)] \leq deg[\varphi_{ii}(L)]$ for $j > i$, or in the final AR equation form where $\Phi(L)$ is scalar. It may be more interesting to impose constraints on the moving average part instead because it is this part which causes problems in the estimation of VARMA models. Other identified representations which do not have a simple MA operator include the reversed echelon canonical form [see Poskitt (1992)] where we the rows of the VARMA model in echelon form are permuted so that the Kronecker indices are ordered from smallest to largest, and the scalar component model [see Tiao and Tsay (1989)] where contemporaneous linear transformations of the vector process are considered. A general treatment of algebraic and topological structure underlying VARMA models is given in Hannan and Kavalieris (1984b).

4. Estimation

We next introduce elements of notation for the parameters of our model. First, irrespective of the VARMA representation employed, we split the whole vector of parameters γ in two parts γ_1 (the parameters for the AR part) and γ_2 (MA part):

$$\gamma = [\gamma_1, \gamma_2]' . \tag{4.1}$$

For a VARMA model in diagonal MA equation form, γ_1 and γ_2 are

$$\gamma_1 = \left[\varphi_{1\bullet,1}, \ldots, \varphi_{1\bullet,p}, \ldots, \varphi_{K\bullet,1}, \ldots, \varphi_{K\bullet,p}\right], \tag{4.2}$$
$$\gamma_2 = \left[\theta_{11,1}, \ldots, \theta_{11,q_1}, \ldots, \theta_{KK,1}, \ldots, \theta_{KK,q_K}\right], \tag{4.3}$$

while for a VARMA model in final MA equation form, γ_2 is

$$\gamma_2 = [\theta_1, \ldots, \theta_q] .$$

For VARMA models in diagonal AR equation form, we simply invert γ_1 and γ_2:

$$\gamma_1 = \left[\varphi_{11,1}, \ldots, \varphi_{11,p_1}, \ldots, \varphi_{KK,1}, \ldots, \varphi_{KK,p_K}\right], \tag{4.4}$$
$$\gamma_2 = \left[\theta_{1\bullet,1}, \ldots, \theta_{1\bullet,q}, \ldots, \theta_{K\bullet,1}, \ldots, \theta_{K\bullet,q}\right], \tag{4.5}$$

while for a VARMA model in final AR equation form,

$$\gamma_1 = \left[\varphi_1, \ldots, \varphi_p\right]. \tag{4.6}$$

The estimation method involves three steps.

Step 1. Estimate a VAR(n_T) to approximate the VARMA(p,q) and recuperate the residuals that we will call $\hat{U}_t$:

$$\hat{U}_t = Y_t - \sum_{l=1}^{n_T} \hat{\Pi}_l^{n_T} Y_{t-l} \tag{4.7}$$

with $T > 2\,K\,n_T$.

Step 2. With the residuals from step 1, compute an estimate of the covariance matrix of U_t, $\hat{\Sigma}_U = \frac{1}{T}\sum_{t=n_T+1}^{T} \hat{U}_t\hat{U}_t'$ and estimate by GLS the multivariate regression

$$\Phi(L)Y_t = [\Theta(L) - I_K]\hat{U}_t + e_t\,, \tag{4.8}$$

to get estimates $\tilde{A}(L)$ and $\tilde{\Theta}(L)$ of $\Phi(L)$ and $\Theta(L)$. The estimator is

$$\tilde{\gamma} = \left[\sum_{t=l}^{T} \hat{Z}_{t-1}' \hat{\Sigma}_U^{-1} \hat{Z}_{t-1}\right]^{-1} \left[\sum_{t=l}^{T} \hat{Z}_{t-1}' \hat{\Sigma}_U^{-1} Y_t\right] \tag{4.9}$$

where $l = n_T + \max(p, q) + 1$. Setting

$$\mathbf{Y}_{t-1}(p) = \left[y_{1,t-1}, \ldots, y_{K,t-1}, \ldots, y_{1,t-p}, \ldots, y_{K,t-p}\right], \tag{4.10}$$
$$\hat{\mathbf{U}}_{t-1} = \left[\hat{u}_{1,t-1}, \ldots, \hat{u}_{K,t-1}, \ldots, \hat{u}_{1,t-q}, \ldots, \hat{u}_{K,t-q}\right], \tag{4.11}$$
$$\mathbf{y}_{k,t-1} = \left[y_{k,t-1}, \ldots, y_{k,t-p_k}\right], \tag{4.12}$$
$$\hat{\mathbf{u}}_{k,t-1} = \left[\hat{u}_{k,t-1}, \ldots, \hat{u}_{k,t-q_k}\right], \tag{4.13}$$

the matrix $\hat{Z}_{t-1}$ for the various representations is:

$$\hat{Z}_{t-1}^{DMA} = \begin{bmatrix} \mathbf{Y}_{t-1}(p) & \cdots & 0 & \hat{\mathbf{u}}_{1,t-1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Y}_{t-1}(p) & 0 & \cdots & \hat{\mathbf{u}}_{K,t-1} \end{bmatrix}, \tag{4.14}$$

$$\hat{Z}_{t-1}^{FMA} = \begin{bmatrix} \mathbf{Y}_{t-1}(p) & \cdots & 0 & \hat{\mathbf{u}}_{1,t-1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \mathbf{Y}_{t-1}(p) & \hat{\mathbf{u}}_{K,t-1} \end{bmatrix}, \tag{4.15}$$

$$\hat{Z}_{t-1}^{DAR} = \begin{bmatrix} \mathbf{y}_{1,t-1} & \cdots & 0 & \hat{\mathbf{U}}_{t-1} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{y}_{K,t-1} & 0 & 0 & \hat{\mathbf{U}}_{t-1} \end{bmatrix}, \tag{4.16}$$

$$\hat{Z}_{t-1}^{FAR} = \begin{bmatrix} \mathbf{y}_{1,t-1} & \hat{\mathbf{U}}_{t-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_{K,t-1} & 0 & 0 & \hat{\mathbf{U}}_{t-1} \end{bmatrix}, \tag{4.17}$$

where DMA, FMA, DAR and FAR respectively stands for Diagonal MA, Final MA, Diagonal AR and Final AR equation form.

Step 3. Using the second step estimates, we first form new residuals

$$\tilde{U}_t = Y_t - \sum_{i=1}^{p} \tilde{\Phi}_i Y_{t-i} + \sum_{j=1}^{q} \tilde{\Theta}_j \tilde{U}_{t-j} \tag{4.18}$$

initiating with $\tilde{U}_t = 0$, $t \leq \max(p,q)$, and we define

$$X_t = \sum_{j=1}^{q} \tilde{\Theta}_j X_{t-j} + Y_t, \tag{4.19}$$

$$W_t = \sum_{j=1}^{q} \tilde{\Theta}_j W_{t-j} + \tilde{U}_t, \tag{4.20}$$

initiating with $X_t = W_t = 0$ for $t \leq \max(p,q)$. We also compute a new estimate of Σ_U, $\tilde{\Sigma}_U = \frac{1}{T}\sum_{t=\max(p,q)+1}^{T} \tilde{U}_t \tilde{U}_t'$. Then we regress by GLS $\tilde{U}_t + X_t - W_t$ on $\tilde{V}_{t-1}$ with

$$\tilde{V}_t = \sum_{j=1}^{q} \tilde{\Theta}_j \tilde{V}_{t-j} + \tilde{Z}_t \tag{4.21}$$

where $\tilde{Z}_t$ is just like $\hat{Z}_t$ from step 2 except that it is computed with $\tilde{U}_t$ instead of $\hat{U}_t$ to obtain regression coefficients that we call $\hat{A}_i$ and $\hat{\Theta}_j$:

$$\hat{\gamma} = \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}' \tilde{\Sigma}_U^{-1} \tilde{V}_{t-1} \right]^{-1} \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}' \tilde{\Sigma}_U^{-1} [\tilde{U}_t + X_t - W_t] \right]. \tag{4.22}$$

The properties of the above estimates are summarized in the following three theorems. Theorem **4.1** is a generalization of results from Lewis and Reinsel (1985) where convergence is demonstrated for mixing rather than i.i.d. innovations. We denote the Euclidean norm by $\|B\|^2 = \mathrm{tr}(B'B)$.

Theorem 4.1 (VARMA first step estimates) *Let (i) the VARMA model be defined by equations* (2.1)*-*(2.6)*; (ii) the strong mixing condition* (2.16) *hold; (iii) assume that* $E[|u_{it}|^{4+2\delta}] < \infty,\ \forall i$

and for some $\delta > 0$. If n_T grows at a rate faster than $\log T$ with ${n_T}^2/T \to 0$, then for the first stage estimates

$$\sum_{l=1}^{n_T} \|\hat{\Pi}_l^{n_T} - \Pi_l\| = O_p(n_T T^{-1/2}). \tag{4.23}$$

Theorem 4.2 (VARMA second step estimates) *Let (i) the VARMA model be defined by equations (2.1)-(2.6) and be identified; (ii) let the strong mixing condition (2.16) hold; (iii) assume that $E[|u_{it}|^{4+2\delta}] < \infty,\ \forall i$ and for some $\delta > 0$. If n_T grows at a rate faster than $\log T$ with ${n_T}^2/T \to 0$, then the second stage estimates converge in quadratic mean to their true value and*

$$\sqrt{T}\,(\tilde{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}\left(0, \tilde{J}^{-1}\tilde{I}\tilde{J}^{-1}\right)$$

where

$$\tilde{I} = \sum_{j=-\infty}^{\infty} \mathsf{E}\left[\left\{Z'_{t-1}\Sigma_U^{-1}U_t\right\}\left\{Z'_{t-1-j}\Sigma_U^{-1}U_{t-j}\right\}'\right], \quad \tilde{J} = \mathsf{E}\left[Z'_{t-1}\Sigma_U^{-1}Z_{t-1}\right],$$

and Z_{t-1} is equal to the matrix $\hat{Z}_{t-1}$ where $\hat{U}_t$ is replaced by U_t. Further, if $m_T^4/T \to 0$ with $m_T \to \infty$ then the matrix $\tilde{I}$ and $\tilde{J}$ can be consistently estimated in probability respectively by

$$\tilde{I}_T = \frac{1}{T}\sum_{j=-m_T}^{m_T} \omega(j, m_T) \sum_{t=l+|j|}^{T} \left\{\hat{Z}'_{t-1}\hat{\Sigma}_U^{-1}\tilde{U}_t\right\}\left\{\hat{Z}'_{t-1-j}\hat{\Sigma}_U^{-1}\tilde{U}_{t-j}\right\}', \tag{4.24}$$

$$\tilde{J}_T = \frac{1}{T}\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}_U^{-1}\hat{Z}_{t-1}, \tag{4.25}$$

with $\omega(j, m_T) = 1 - |j|/(m_T + 1)$.

Theorem 4.3 (VARMA third step estimates) *Let (i) the VARMA model be defined by equations (2.1)-(2.6) and be identified; (ii) let the strong mixing condition (2.16) hold; (iii) assume that $E[|u_{it}|^{4+2\delta}] < \infty,\ \forall i$ and for some $\delta > 0$. If n_T grows at a rate faster than $\log T$ with ${n_T}^2/T \to 0$, then the third stage estimates converge in quadratic mean to their true value, and*

$$\sqrt{T}\,(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}\left(0, \hat{J}^{-1}\hat{I}\hat{J}^{-1}\right) \tag{4.26}$$

with

$$\hat{I} = \sum_{j=-\infty}^{\infty} \mathsf{E}\left[\left\{V'_{t-1}\Sigma_U^{-1}U_t\right\}\left\{V'_{t-1-j}\Sigma_U^{-1}U_{t-j}\right\}'\right], \quad \hat{J} = \mathsf{E}\left[V'_{t-1}\Sigma_U^{-1}V_{t-1}\right] \tag{4.27}$$

and V_{t-1} is equal to the matrix $\tilde{V}_{t-1}$ where $\tilde{U}_t$ is replaced by U_t. Further, if $m_T^4/T \to 0$ with

$m_T \to \infty$ *then the matrix* $\hat{I}$ *and* $\hat{J}$ *can be consistently estimated in probability respectively by*

$$\hat{I}_T = \frac{1}{T} \sum_{j=-m_T}^{m_T} \omega(j, m_T) \sum_{t=l'+|j|}^{T} \left\{ \tilde{V}'_{t-1} \tilde{\Sigma}_U^{-1} \bar{U}_t \right\} \left\{ \tilde{V}'_{t-1-j} \tilde{\Sigma}_U^{-1} \bar{U}_{t-j} \right\}' \quad (4.28)$$

$$\hat{J}_T = \frac{1}{T} \sum_{t=\max(p,q)+1}^{T} \tilde{V}'_{t-1} \tilde{\Sigma}_U^{-1} \tilde{V}_{t-1} \quad (4.29)$$

with $l' = \max(p, q) + 1$ *and* $\bar{U}_t$ *are the filtered residuals computed with* $\hat{\gamma}$.

Notice the simplicity of this estimation method. Only three regressions are needed so we can avoid all the caveats associated with nonlinear optimizations. This is an important problem with VARMA models where one typically deals with a high number of parameters and numerical convergence may be hard to obtain. This is especially important when we consider the fact that the asymptotic distribution of our estimators, on which we would base our inference, may be a bad approximation to the finite-sample distribution in high-dimensional dynamic models. Because of this, an estimation procedure which only requires linear methods is interesting since it suggests that simulation-based procedures – bootstrap techniques for example – should be used, something that would be impractical if the estimation is based on non-linear optimizations.

It is also important to mention that this procedure is not specific to the representations considered in this work. The expressions can be easily adapted to other identified representation, *e.g.* the echelon form. Since our estimation method is only based on regressions we can afford to use a less parsimonious representation whereas for nonlinear method it is highly important to keep the number of parameters to a minimum.

For the estimation of VARMA models the emphasis has been on maximizing the likelihood (minimizing by nonlinear least squares) quickly. There are two ways of doing this. The first is having quick and efficient algorithm to evaluate the likelihood [*e.g.* Luceño (1994) and the reference therein, Mauricio (2002), Shea (1989)]. The second is to find preliminary consistent estimates that can be computed quickly to initialize the optimization algorithm. We are not the first to present a generalization to VARMA models of the Hannan and Rissanen (1982) estimation procedure for ARMA models [whose asymptotic properties are further studied in Zhao-Guo (1985) and Saikkonen (1986)]; see also Durbin (1960), Hannan and Kavalieris (1984a), Hannan, Kavalieris, and Mackisack (1986), Poskitt (1987), Koreisha and Pukkila (1990a, 1990b, 1995), Pukkila, Koreisha, and Kallinen (1990), Galbraith and Zinde-Walsh (1994, 1997). A similar method in three steps is also presented in Hannan and Kavalieris (1984a) where the third step is presented as a correction to the second step estimates.

There are many variations around the innovation-substitution approach for the estimation of VARMA models but with the exception of Hannan and Kavalieris (1984b) and us, none use a third step to improve the efficiency of the estimators, surely because these procedures are often seen as a way to get initial values to start up a nonlinear optimization [*e.g.*, see Poskitt (1992), Koreisha and Pukkila (1989), Lütkepohl and Claessen (1997)]. In one of them, Koreisha and Pukkila (1989), lagged and current innovations are replaced by the corresponding residuals and a regression is performed. This is asymptotically the same as the first two steps of our method. Other variations are described in Hannan and Kavalieris (1986), Hannan and Deistler (1988), Huang and Guo (1990), Spliid (1983), Reinsel, Basu, and Yap (1992), Poskitt and Lütkepohl (1995), Lütkepohl and Poskitt (1996b) and Flores de Frutos and Serrano (2002). Another approach is to use the link that exist between the VARMA parameters and the infinite VAR and VMA representations. See Galbraith, Ullah, and Zinde-Walsh (2000) for the estimation of VMA models using a VAR.

Here, however, we supply a distributional theory which holds under much weaker assumptions. In the articles cited above, the data generating processes considered have innovations that are either i.i.d. or at a minimum form a martingale difference sequence. This allow us to study a broader class of models, *e.g.* temporally aggregated processes, marginalized processes, weak representation of nonlinear models.

We can ask ourselves what is the cost of not doing the nonlinear estimation. For a given sample size we will certainly lose some efficiency because of the first step estimation. We can nonetheless compare the asymptotic variance matrix of our estimates with the corresponding nonlinear estimates. We first can see that if the innovations are a m.d.s., then the asymptotic variance of our linear estimates is the same as the variance of maximum likelihood estimates under Gaussianity. The variance of maximum likelihood estimates for i.i.d. Gaussian innovations is given in Lütkepohl (1993):

$$I = plim \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial U_t'}{\partial \gamma} \Sigma^{-1} \frac{\partial U_t}{\partial \gamma'} \right]^{-1} . \tag{4.30}$$

We can transform this expression so as to obtain an equation more closely related to our previous results. First, we split γ in the same two vectors γ_1 (the AR parameters) and γ_2 (the MA parameters), then we compute $\partial U_t / \partial \gamma_1'$ and $\partial U_t / \partial \gamma_2'$. We know that

$$U_t = Y_t - \Phi_1 Y_{t-1} - \cdots - \Phi_p Y_{t-p} + \Theta_1 U_{t-1} + \cdots + \Theta_q U_{t-q}. \tag{4.31}$$

So taking the derivative with respect to γ_1':

$$\begin{aligned} \frac{\partial U_t}{\partial \gamma_1'} &= Z_{\bullet 1:dim(\gamma_1),t-1} + \Theta_1 \frac{\partial U_{t-1}}{\partial \gamma_1'} + \cdots + \Theta_q \frac{\partial U_{t-q}}{\partial \gamma_1'} , && (4.32) \\ \Theta(L) \frac{\partial U_t}{\partial \gamma_1'} &= Z_{\bullet 1:dim(\gamma_1),t-1} , && (4.33) \\ \frac{\partial U_t}{\partial \gamma_1'} &= \Theta(L)^{-1} Z_{\bullet 1:dim(\gamma_1),t-1} , && (4.34) \end{aligned}$$

where $Z_{\bullet 1:dim(\gamma_1),t-1}$ is the first $\dim(\gamma_1)$ columns of Z_{t-1}. Similarly the derivative with respect to γ_2' is

$$\begin{aligned} \frac{\partial U_t}{\partial \gamma_2'} &= Z_{\bullet dim(\gamma_1)+1:dim(\gamma),t-1} + \Theta_1 \frac{\partial U_{t-1}}{\partial \gamma_2'} + \cdots + \Theta_q \frac{\partial U_{t-q}}{\partial \gamma_2'} \\ &= \Theta(L)^{-1} Z_{\bullet dim(\gamma_1)+1:dim(\gamma),t-1} && (4.35) \end{aligned}$$

Combining the two expressions we see that

$$\frac{\partial U_t}{\partial \gamma'} = V_{t-1} \tag{4.36}$$

so the variance matrix for maximum likelihood estimates I is equal to the matrix J^{-1} from the third step estimation. Moreover if U_t is a m.d.s. we see that we have the equality $J = I$ so that the asymptotic variance matrix that we get in the third step of our method is the same as one would get by doing the maximum likelihood.

If the innovations are merely uncorrelated, then we can generalize the results of Francq and Zakoïan (1998) who proved the consistency of nonlinear least squares for univariate weak ARMA models. The authors show that the asymptotic distribution of the estimates are

$$\sqrt{T}\,(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}\left(0, J^{-1} I J^{-1}\right) \tag{4.37}$$

where $\gamma = \left\{\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q\right\}$,

$$I = 4 \sum_{k=-\infty}^{\infty} Cov\left[u_t \frac{\partial u_t}{\partial \gamma} \; ; \; u_{t-k} \frac{\partial u_{t-k}}{\partial \gamma}\right] \;, \quad J = 2E\left[\frac{\partial u_t}{\partial \gamma} \, \frac{\partial u_t}{\partial \gamma'}\right]. \tag{4.38}$$

Without formally proving it we can generalize these expressions for the multivariate case. Writing the multivariate nonlinear least squares problem and doing a first order expansion of the first order condition we find that the expression for the asymptotic covariance matrix of the estimates would again be $J^{-1} I J^{-1}$ with

$$I = 4 \sum_{k=-\infty}^{\infty} Cov\left[U_t \, \Sigma^{-1} \, \frac{\partial U_t}{\partial \gamma'} \; ; \; U_{t-k} \, \Sigma^{-1} \, \frac{\partial U_{t-k}}{\partial \gamma'}\right] \;, \quad J = 2E\left[\frac{\partial U_t'}{\partial \gamma} \, \Sigma^{-1} \, \frac{\partial U_t}{\partial \gamma'}\right] \tag{4.39}$$

In our previous results we saw that $\partial U_t / \partial \gamma' = V_{t-1}$. From this we see that $J = 2\hat{J}$, $I = 4\hat{I}$ and our third-step estimator have the same asymptotic variance-covariance matrix as maximum likelihood or non-linear least squares estimators depending on the properties of the innovations. To get a feel for the loss of efficiency in finite samples due to replacing the true innovations by residuals from a long VAR we performed Monte Carlo simulations and report the results in section 6.

5. Order selection

We still have unknowns in our model, the orders of the AR and MA operators. If no theory specifies these parameters, we have to use a statistical procedure to choose them. We propose the following information criterion method to choose the orders for VARMA models in the different identified representations proposed in Section 3. In the second step of the estimation, we compute for all

$p_i \leq P$ and $q_i \leq Q$ the following information criterion:

$$\log(\det \tilde{\Sigma}_U) + \dim(\gamma)\frac{(\log T)^{1+\delta}}{T}, \quad \delta > 0. \tag{5.1}$$

We then choose $\hat{p}_i$ and $\hat{q}_i$ as the set which minimizes the information criterion. We assume that the upper bound P and Q on the order of the AR and MA part are bigger than the true values of p_i and q_i (or that they slowly grow with the sample size). The properties of $\hat{p}_i$ and $\hat{q}_i$ are summarized in the following theorem.

Theorem 5.1 (Estimation of the order p and q in VARMA models) *Let (i) the VARMA model be defined by equations (2.1)-(2.6) and be identified; (ii) let the strong mixing condition (2.16) hold; (iii) assume that $E[|u_{it}|^{4+2\delta}] < \infty,\ \forall i$ and for some $\delta > 0$. If n_T grows at a rate faster than $\log T$ with ${n_T}^2/T \to 0$ and the orders are chosen according to (5.1), then $\hat{p}_i$ and $\hat{q}_i$, $i = 1, \ldots, K$, converge in probability to their true value.*

In practice, this procedure can lead to a search over too many models for the diagonal representations. A valid alternative is to search for the true orders by proceeding equation by equation. In the second step of the estimation, instead of doing a simultaneous estimation, just perform univariate regressions. For a VARMA model in diagonal MA equation form, regress

$$y_{kt} = \sum_{i=1}^{p_k}\sum_{j=1}^{K}\varphi_{kj,i}y_{j,t-i} - \sum_{j=1}^{q_k}\theta_{kk,j}\hat{u}_{k,t-j} + e_{kt}, \tag{5.2}$$

for $k = 1, \ldots, K$, while for a VARMA models in diagonal AR equation form we regress

$$y_{kt} = \sum_{i=1}^{p_k}\varphi_{kk,i}y_{k,t-i} - \sum_{i=1}^{q_k}\sum_{j=1}^{K}\theta_{kj,i}\hat{u}_{j,t-i} + e_{kt}. \tag{5.3}$$

We then chose $\hat{p}_k$ and $\hat{q}_k$ as the orders which minimize the following information criterion:

$$\log(\sigma_k^2) + g(p_k, q_k)\frac{(\log T)^{1+\delta}}{T} \tag{5.4}$$

where $\delta > 0$ and $g(p_k, q_k) = p_k K + q_k$ or $g(p_k, q_k) = p_k + q_k K$ for the diagonal MA or AR equation form representation respectively. The global order for the autoregressive operator is then $\hat{p} = \max(\hat{p}_1, \ldots, \hat{p}_K)$ for the diagonal MA representation and, similarly for the diagonal AR representation, $\hat{q} = \max(\hat{q}_1, \ldots, \hat{q}_K)$. We see that this equation by equation selection procedure is not only easier to apply, it can lead to more parsimonious representations by identifying rows of zeros coefficients in Φ_i or Θ_j.

Theorem 5.2 (Estimation of the order p and q in diagonal VARMA models) *Let (i) the VARMA model be defined by equations (2.1)-(2.6) and be in either the diagonal MA or AR equation form; (ii) let the strong mixing condition (2.16) hold; (iii) assume that $E[|u_{it}|^{4+2\delta}] < \infty,\ \forall i$ and for* *some $\delta > 0$. If n_T grows at a rate faster than $\log T$ with ${n_T}^2/T \to 0$ and the orders are chosen according to (5.4), then $\hat{p}_i$ and $\hat{q}_i$, $i = 1, \ldots, K$, converge in probability to their true value.*

The criterion in equation (5.1) is a generalization of the information criterion proposed by Hannan and Rissanen (1982) which the authors acknowledged that it must in fact be modified to provide consistent estimates of the order, p and q. The original criterion was

$$\log \tilde{\sigma}^2 + (p+q)\frac{(\log T)^\delta}{T} \tag{5.5}$$

with $\delta > 0$. But in Hannan and Rissanen (1983) they acknowledged that $\tilde{\sigma}^2 - \sigma^2$ is $O(n_T T^{-1})$ and not $O(T^{-1})$ so the penalty $(\log T)^\delta/T$ is not strong enough. The authors proposed two possible modifications to their procedure. The simpler is to take $(\log T)^{1+\delta}$ instead of $(\log T)^\delta$ in the information criterion so that the penalty on $p+q$ will dominate $\log \hat{\sigma}^2$ in the criterion. The second, which they favored and was used in later work [see Hannan and Kavalieris (1984b)], is to modify the first step of the procedure. Instead of taking $n_T = O(\log T)$ they used another information criterion to choose the order of the long autoregression and they iterated the whole procedure picking a potentially different p and q at every iteration. A similar approach is also proposed in Poskitt (1987). In this work we prefer the first solution so as to keep the procedure as simple as possible.

For the identification of the order of VARMA models, it all depends on the representation used. Although it was one of the first representation studied, not much work has been done with the final AR equation form. People felt that this representation gives VARMA models with too many parameters. A complete procedure to fit VARMA models under this representation is given in Lütkepohl (1993): One would first fit an ARMA(p_i, q_i) model to every univariate time series, using maybe the procedure of Hannan and Rissanen (1982). To build the VARMA representation in final AR equation form, knowing that the VAR operator is the same for every equation we would take it to be the product of all the univariate AR polynomials. This would give a VAR operator of order $p = \sum_{i=1}^{K} p_i$. Accordingly, for the VMA part we would take $q = \max_k[q_k + \sum_{i=1, i\neq k}^{K} p_i]$. It is no wonder that people feel that the final equation form uses to many parameters.

For VARMA models in echelon form, there has been a lot more work done on the identification of Kronecker indices. The problem has been studied by, among others, Hannan and Kavalieris (1984b), Poskitt (1992) and Lütkepohl and Poskitt (1996b). Non-stationary or cointegrated systems are considered by Huang and Guo (1990), Bartel and Lütkepohl (1998), and Lütkepohl and Claessen (1997). Additional references are given in Lütkepohl (1993, Chapter 8).

A complementing approach to specify VARMA models, which is based on Cooper and Wood (1982), aims at finding simplifying structures via some combinations of the different series to obtain more parsimonious models. It includes Tiao and Tsay (1989), Tsay (1989a, 1989b, 1991) and Nsiri and Roy (1992, 1996).

The final stage of ARMA model specification usually involve analyzing the residuals, *i.e.* checking if they are uncorrelated. Popular tools include portmanteau tests such as Box-Pierce [Box and Pierce (1970)] and Ljung-Box [Ljung and Box (1978)] tests, and their multivariate generalization [Lütkepohl (1993, Section 5.2.9)]. Those tests are not directly applicable in our case because they are derived under strong assumptions for the innovations (independence or martingale difference).

Fortunately, we could extend to multivariate ARMA models the theory for BP and LB tests developed by Francq, Roy, and Zakoïan (2003) for weak ARMA models. Such extension is left as future work.

6. Monte Carlo simulations

To illustrate the performance of our estimation method we ran two types of simulations. For the first type, weak VARMA models were simulated where the innovations are not independent nor a m.d.s. but merely uncorrelated. The second type of simulations involves strong VARMA models (VARMA models with i.i.d. Gaussian innovations). All the simulated models are bivariate so the results are easier to analyze. The results are generated using Ox version 3.30 on Linux [see Doornik (1999)]. We performed 1000 simulations for each model. The results with strong VARMA models being comparable to those for weak VARMA models, we only report results for the latter.

We simulate weak VARMA processes by directly simulating weak innovations, from which we build the simulated series. From the results in Drost and Nijman (1993), we know that the temporal aggregation of a strong GARCH process (where the standardized innovations are i.i.d.) will give a weak process[1]. Suppose $\tilde{U}_t$ is given by the following bivariate ARCH model:

$$\tilde{U}_t = H_t^{1/2}\varepsilon_t \ , \quad H_t = \Omega + \alpha\tilde{U}_{t-1}\tilde{U}_{t-1}' \tag{6.1}$$

where ε_t is i.i.d. $N(0, I_2)$, $H_t^{1/2}$ is the Cholesky decomposition of H_t and α is a scalar. If we consider $\tilde{U}_t$ as a stock variable, then temporal aggregation of $\tilde{U}_t$ over two periods, *i.e.*

$$U_t = \tilde{U}_{2t} \tag{6.2}$$

will give a weak process. The series U_t will be uncorrelated but not a m.d.s., its mean will be zero and the variance will be $\Omega(1-\alpha^2)/(1-\alpha)$.

In these examples, because the innovations are not a m.d.s., we cannot do maximum likelihood. We instead employ nonlinear generalized least-squares (GLS), *i.e.* we minimize the nonlinear least squares, compute an estimate of the variance matrix of the innovations and then do nonlinear GLS. We did not apply this procedure, partly to reduce the estimation time in our Monte Carlo study, partly because there is no asymptotic gain in iterating.

In these simulations the sample size is 250 observations, which represent about 20 years of monthly data, a reasonable sample size for macroeconomic data. Tables 1 gives results for a VARMA model in final MA equation form [VARMA$(1,1)$], while results for VARMA models in diagonal MA equation form are given in Tables 2 and 3 [VARMA$(1,1)$ with $q = (1,1)$ and VARMA(2,1) with $q = (1,1)$ respectively]. We present the results (mean, standard deviations, root mean square error, 5% quantile, 95% quantile and median) for the second (when the number of parameters does not exceed five) and third step estimates, and the nonlinear GLS estimates (using

[1]Another way of simulating a weak VARMA process is to time-aggregate a strong VARMA process with innovations that have skewed marginal distributions (*e.g.*, a mixture of two Gaussian distributions with different means but mean zero unconditionally). We can appeal to the results of Francq and Zakoïan (1998, Section 2.2.1) to claim that the resulting VARMA is only weak.

the true value of the parameters as initial values). Samples for which the optimization algorithm did not converge were dropped (this happened for less than 1% of the simulations). In our simulations, we took

$$\Omega = \begin{bmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{bmatrix}, \quad \alpha = 0.3. \tag{6.3}$$

From looking at the RMSE, a first thing to notice is that there can be sizable improvement in doing the third step. Some of the third step RMSEs in Tables 1 and 2 are more than 50% smaller than for the second step. This is an interesting observation considering that the third step basically involve only one extra regression. Comparing the third step RMSEs and the RMSEs for the nonlinear GLS estimates, we see that the former are usually no more than 15% bigger. This is also an interesting observation. The cost of avoiding a numerical optimization, which can become quite challenging as the number of time series studied or order of the operators increases, appears to be small.

In the top part of these tables we also present the results for the selection of the order of the operators using our proposed information criterion. For models in final MA equation form, we have to select the orders p and q, and for models in diagonal MA equation, the selection is over p, q_1 and q_2. In Table 1, we see that for VARMA models in final MA equation form the most frequently chosen orders are the true ones, and the criterion will tend to pick a higher value for q than for p. This result might partially be skewed by the fact that the simulated models have a highly persistent moving average ($\theta_1 = 0.9$). For VARMA models in diagonal equation form (Tables 2 and 3), we get similar results. The orders which are selected with the highest frequency are the true ones, but for some models we pick the wrong orders more than 50% of the time.

These results for the information criterion are fairly sensitive to the value of δ and c_0, more so for the model with a diagonal representation. This can be compared to non-parametric regressions and the selection of the bandwidth parameter. The performance of the information criterion with respect to these two parameters should be investigated further.

7. Application to a macroeconomics model of the U.S. monetary policy

To illustrate our estimation method and the gains that can be obtained from using a more parsimonious representation, we fit VARMA and VAR models to six macroeconomic time series and compute the impulse-response functions generated by each model. What people typically do to get the impulse-response functions is first fit a VAR to their multiple time series and then get the implied infinite VMA representation. The order of the VAR required for macro series is usually high. For example, Bernanke and Mihov (1998) use a VAR(13) to model six monthly macroeconomic time series when about 30 years of data are available. The resulting standard errors for the impulse-response functions are very large, like in most macroeconomic study. We can ask ourselves how much of this is due to the fact that so many parameters are estimated. To try to answer this we will study the impulse-response functions generated by VARMA models estimated on the same data. We will concentrate on VARMA models in final MA equation form.

Our example is based on McMillin (2001) who compare numerous identification restrictions

Table 1. Estimation of a weak final MA equation form VARMA$(1,1)$.
Weak final MA equation form VARMA$(1,1)$.

The simulated model is a weak VARMA$(1,1)$ in final MA equation form with $\varphi_{11,1}=0.5$, $\varphi_{12,1}=-0.6$, $\varphi_{21,1}=0.7$, $\varphi_{22,1}=0.3$ and $\theta_1=0.9$. The variance of the innovations is 1.3 and the covariance is 0.91. Sample size is 250, the length of the long AR is $n_T=20$, the number of repetition is 1000. The parameter in the criterion is $\delta=0.3$.

$p \setminus q$	0	1	2	3	4	5
0	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.736	0.101	0.024	0.007	0.003
2	0.000	0.002	0.107	0.013	0.003	0.003
3	0.000	0.000	0.000	0.001	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Second step							
$a_1(1,1)$	0.5	0.4255	0.0629	0.0975	0.3243	0.5307	0.4281
$a_1(1,2)$	-0.6	-0.6390	0.0515	0.0646	-0.7239	-0.5541	-0.6401
$a_1(2,1)$	0.7	0.6682	0.0586	0.0666	0.5677	0.7599	0.6686
$a_1(2,2)$	0.3	0.2117	0.0551	0.1041	0.1195	0.3043	0.2129
b_1	0.9	0.8128	0.0593	0.1054	0.7148	0.9079	0.8139
Third step							
$a_1(1,1)$	0.5	0.5001	0.0505	0.0505	0.4174	0.5857	0.5006
$a_1(1,2)$	-0.6	-0.5896	0.0469	0.0481	-0.6685	-0.5154	-0.5899
$a_1(2,1)$	0.7	0.6859	0.0524	0.0543	0.6018	0.7682	0.6852
$a_1(2,2)$	0.3	0.3111	0.0494	0.0507	0.2341	0.3911	0.3101
b_1	0.9	0.8978	0.0348	0.0349	0.8368	0.9494	0.9000
NLLS							
$a_1(1,1)$	0.5	0.4952	0.0504	0.0507	0.4120	0.5789	0.4962
$a_1(1,2)$	-0.6	-0.6089	0.0432	0.0441	-0.6813	-0.5402	-0.6094
$a_1(2,1)$	0.7	0.7017	0.0494	0.0494	0.6209	0.7810	0.7023
$a_1(2,2)$	0.3	0.2875	0.0460	0.0476	0.2138	0.3660	0.2868
b_1	0.9	0.8866	0.0282	0.0312	0.8378	0.9294	0.8884

Table 2. Estimation of a weak diagonal MA equation form VARMA$(1,1)$
Weak diagonal MA equation form VARMA$(1,1)$

The simulated model is a weak VARMA$(1,1)$ in diagonal MA equation form with $\varphi_{11,1} = 0.5$, $\varphi_{12,1} = -0.6$, $\varphi_{21,1} = 0.7$, $\varphi_{22,1} = 0.3$, $\theta_{1,1} = 0.9$ and $\theta_{1,1} = 0.7$. The variance of the innovations is 1.3 and the covariance is 0.91. Sample size is 250, the length of the long AR is $n_T = 20$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.3$.

(p, q_1, q_2)	Frequency	(p, q_1, q_2)	Frequency
1,1,1	0.588	1,1,3	0.026
1,2,1	0.123	2,1,1	0.014
1,1,2	0.062	1,4,1	0.014
1,3,1	0.045	1,5,1	0.012
2,2,2	0.043	1,1,5	0.010

	Value	Average	Std. dev.	RMSE	5%	95%	Median
Second step							
$a_1(1,1)$	0.5	0.4277	0.0601	0.0940	0.3284	0.5233	0.4303
$a_1(1,2)$	-0.6	-0.6439	0.0507	0.0671	-0.7291	-0.5594	-0.6444
$a_1(2,1)$	0.7	0.6732	0.0514	0.0579	0.5863	0.7550	0.6729
$a_1(2,2)$	0.3	0.2314	0.0526	0.0865	0.1446	0.3193	0.2309
$b_1(1)$	0.9	0.8130	0.0707	0.1122	0.6976	0.9266	0.8150
$b_1(2)$	0.7	0.6364	0.0708	0.0952	0.5185	0.7476	0.6393
Third step							
$a_1(1,1)$	0.5	0.5064	0.0469	0.0473	0.4324	0.5845	0.5062
$a_1(1,2)$	-0.6	-0.5960	0.0552	0.0554	-0.6762	-0.5183	-0.5969
$a_1(2,1)$	0.7	0.6988	0.0418	0.0418	0.6314	0.7659	0.6997
$a_1(2,2)$	0.3	0.3021	0.0469	0.0469	0.2272	0.3830	0.3032
$b_1(1)$	0.9	0.8885	0.0442	0.0456	0.8100	0.9531	0.8910
$b_1(2)$	0.7	0.6967	0.0522	0.0523	0.6092	0.7843	0.6969
NLLS							
$a_1(1,1)$	0.5	0.4973	0.0453	0.0453	0.4222	0.5703	0.4972
$a_1(1,2)$	-0.6	-0.6116	0.0443	0.0458	-0.6864	-0.5371	-0.6114
$a_1(2,1)$	0.7	0.7009	0.0411	0.0411	0.6334	0.7683	0.7006
$a_1(2,2)$	0.3	0.2897	0.0441	0.0453	0.2185	0.3645	0.2893
$b_1(1)$	0.9	0.8874	0.0349	0.0371	0.8260	0.9385	0.8894
$b_1(2)$	0.7	0.6950	0.0446	0.0449	0.6198	0.7673	0.6955

Table 3. Estimation of a weak diagonal MA equation form VARMA(2,1)
Weak diagonal MA equation form weak VARMA(2,1).

The simulated model is a weak VARMA(2,1) in diagonal MA equation form with $\varphi_{11,1} = 0.9$, $\varphi_{12,1} = -0.5$, $\varphi_{21,1} = 0.3$, $\varphi_{22,1} = 0.1$, $\varphi_{11,2} = -0.1$, $\varphi_{12,2} = -0.2$, $\varphi_{21,2} = 0.1$, $\varphi_{22,2} = -0.15$, $\varphi_{1,1} = 0.9$, and $\varphi_{2,1} = 0.7$. The variance of the innovations is 1.3 and the covariance is 0.91. Sample size is 250, the length of the long AR is $n_T = 20$, the number of repetition is 1000. The parameter in the criterion is $\delta = 0.2$.

(p, q_1, q_2)	Frequency	(p, q_1, q_2)	Frequency
2,1,1	0.267	2,1,0	0.047
1,2,1	0.204	2,2,1	0.031
2,3,0	0.057	2,4,0	0.029
1,3,1	0.051	2,2,2	0.020
2,3,1	0.050	0,3,1	0.020

	Value	Average	Std. dev.	RMSE	5%	95%	Median
third step							
$a_1(1,1)$	0.90	0.9205	0.1016	0.1036	0.7554	1.0882	0.9204
$a_1(1,2)$	-0.50	-0.5137	0.0922	0.0932	-0.6643	-0.3620	-0.5142
$a_1(2,1)$	0.30	0.3036	0.0802	0.0802	0.1737	0.4326	0.3020
$a_1(2,2)$	0.10	0.1071	0.1666	0.1668	-0.1533	0.3802	0.1037
$a_2(1,1)$	-0.10	-0.0716	0.0937	0.0979	-0.2302	0.0781	-0.0715
$a_2(1,2)$	-0.20	-0.1976	0.1262	0.1262	-0.3995	0.0159	-0.1995
$a_2(2,1)$	0.10	0.1014	0.1127	0.1127	-0.0969	0.2749	0.1111
$a_2(2,2)$	-0.15	-0.1326	0.1363	0.1374	-0.3462	0.1156	-0.1440
$b_1(1)$	0.90	0.8917	0.0774	0.0778	0.7654	1.0079	0.8973
$b_1(2)$	0.70	0.7084	0.1423	0.1426	0.4724	0.9397	0.7112
NLLS							
$a_1(1,1)$	0.90	0.8787	0.0914	0.0939	0.7254	1.0347	0.8799
$a_1(1,2)$	-0.50	-0.5015	0.0918	0.0918	-0.6523	-0.3517	-0.5008
$a_1(2,1)$	0.30	0.2957	0.0801	0.0802	0.1665	0.4244	0.2927
$a_1(2,2)$	0.10	0.0715	0.1252	0.1284	-0.1412	0.2643	0.0748
$a_2(1,1)$	-0.10	-0.0815	0.0887	0.0906	-0.2250	0.0622	-0.0814
$a_2(1,2)$	-0.20	-0.2328	0.1144	0.1190	-0.4268	-0.0491	-0.2332
$a_2(2,1)$	0.10	0.1243	0.0901	0.0933	-0.0216	0.2639	0.1278
$a_2(2,2)$	-0.15	-0.1831	0.1074	0.1124	-0.3522	-0.0050	-0.1826
$b_1(1)$	0.90	0.8861	0.0404	0.0427	0.8133	0.9448	0.8892
$b_1(2)$	0.70	0.6789	0.0883	0.0908	0.5215	0.8065	0.6889

for the structural effects of monetary policy shocks using the same dataset as Bernanke and Mihov (1998).[2] The series are plotted in Figure 1. One of the model studied is a VAR applied to the first difference of the series, in order, *gdpm*, (*psscom-pgdpm*), *fyff*, *nbrec1*, *tr1*, *psscom*. With an argument based on Keating (2002), the author state that using this ordering of the variables the Cholesky decomposition, based on long-run macroeconomic restrictions, which are described in an appendix, of the variance matrix of the innovations will identify the structural effects of the policy variable *nbrec1* without imposing any contemporaneous restrictions among the variables. Since the model is in first difference, the impulse-response at a given order is the cumulative shocks up to that order.

By fitting a VAR(12) to these series we get basically the same impulse-response functions and confidence bands as in McMillin (2001) They are plotted in Figure 2. The impulse-response function for the output and federal funds rate tends to zero as the order increases which is consistent with the notion that a monetary variable does not have a long term impact on real variables. The impulse response of the price level increases as we let the order grow and does not revert to zero.

We next estimate VARMA models for the four representations proposed in this work. The information criterion picked a VARMA(3,10) for the final MA representation. The impulse-response functions for this model are plotted in Figure 3. The behavior of the impulse-response function for GDP, the federal funds rate and the price level from the VARMA models are similar to what we obtained with a VAR. The most notable differences are that the initial decrease in the federal funds rate is smaller (0.20 versus 0.32 percentage point) and the GDP is peaking earlier with the VARMA.

It is not surprising that VAR and VARMA models are giving similar impulse-response functions since they both are a way of getting an infinite MA representation. What is more interesting is the comparison of the width of the confidence bands for the VAR and VARMA's impulse-response functions.[3] For GDP and the federal funds rate, we see that the bands are much smaller for the VARMA model and they shrink more quickly as the horizon increases. The confidence bands for these two variables should be collapsing around their IRF since there should be no long-term effect of the policy variable so the uncertainty should decrease as the horizon increases. The situation is different for the price level. For this variable the confidence band grows with the order. Again this is not so surprising because we expect that a change in the non-borrowed reserves should have a long-term impact on the price level. With a non-dying impact it is natural that the uncertainty about this impact can grow as time passes.

The result that the confidence bands around IRFs can be shorter with a VARMA than with a VAR could be expected since these models are simple extensions of the VAR approach. The introduction of a simple MA operator allows the reduction of the required AR order so we can get more precise estimates, which translate into more precise impulse-response functions.

Another way of comparing the performance of VAR and VARMA models is to compare their

[2]The dataset consist of the log of the real GDP (*gdpm*), total bank reserves (*tr1*), nonborrowed reserves (*nbrec1*), federal funds rate (*fyff*), log of the GDP deflator (*pgdpm*), log of the Dow-Jones index of spot commodity prices (*psccom*). These are monthly data and cover the period January 1962 to December 1996. The monthly data for real GDP and the GDP deflator were constructed by state space methods, using a list of monthly interpolator variables and assuming that the interpolation error is describable as an AR(1) process. Both total reserves and nonborrowed reserves are normalized by a 36-month moving average of total reserves.

[3]The confidence bands are computed by performing a parametric bootstrap using Gaussian innovations.

Table 4. RMSE for VAR and VARMA models]RMSE for VAR and VARMA models

Step ahead	VAR	VARMA diag. MA	VARMA final MA
1	0.0834	0.0764	0.0743
	$p = 1$	$p = 0$	$p = 0$
		$q = (1, 2, 2, 1, 1, 1)$	$q = 12$
3	0.0799	0.0788	0.0744
	$p = 1$	$p = 1$	$p = 1$
		$q = (1, 1, 1, 1, 0, 1)$	$q = 12$
6	0.0826	0.0767	0.0790
	$p = 7$	$p = 3$	$p = 1$
		$q = (4, 4, 1, 4, 0, 4)$	$q = 12$
9	0.0871	0.0774	0.0829
	$p = 2$	$p = 4$	$p = 0$
		$q = (5, 5, 3, 4, 5, 5)$	$q = 12$
12	0.0819	0.0728	0.0803
	$p = 4$	$p = 4$	$p = 1$
		$q = (3, 5, 3, 4, 5, 5)$	$q = 12$

out-of-sample forecasts using a metric (*e.g.*, RMSE as in our example). Employing the same dataset as above, we recursively estimated the models and computed the out-of-sample forecasts, starting at observation 300 until the end of the sample. The orders of the different models are chosen by minimizing the RMSE over the possible values (we impose an upper bound)[4]. The results for the VAR, VARMA diagonal MA and VARMA final MA representations are presented in Table 4. We see that reduction of up to 11% of the RMSE can be obtained by using a VARMA model instead of a VAR.

8. Conclusion

In this paper, we proposed a modeling and estimation method which ease the use of VARMA models. We first propose new identified VARMA representations, the final MA equation form and the diagonal MA equation form. These two representations are simple extensions of the class of VAR models where we add a simple MA operator, either a scalar or a diagonal operator. The addition of a MA part can give more parsimonious representations, yet the simple form of the MA operators does not introduce undue complications.

To ease the estimation we studied the problem of estimating VARMA models by relatively simple methods which only require linear regressions. For that purpose, we considered a generalization

[4]For the VARMA diagonal MA representation we don't search over all the possible orders because it would involve the estimation of too many models. We instead proceed in two steps. We first impose that all the q_i orders are equal which gives us an upper bound for the value of MA orders. In a second step, one equation after the other we check to see if a lower order for the MA order of that equation would lower the RMSE.

Figure 1. Macroeconomic series.

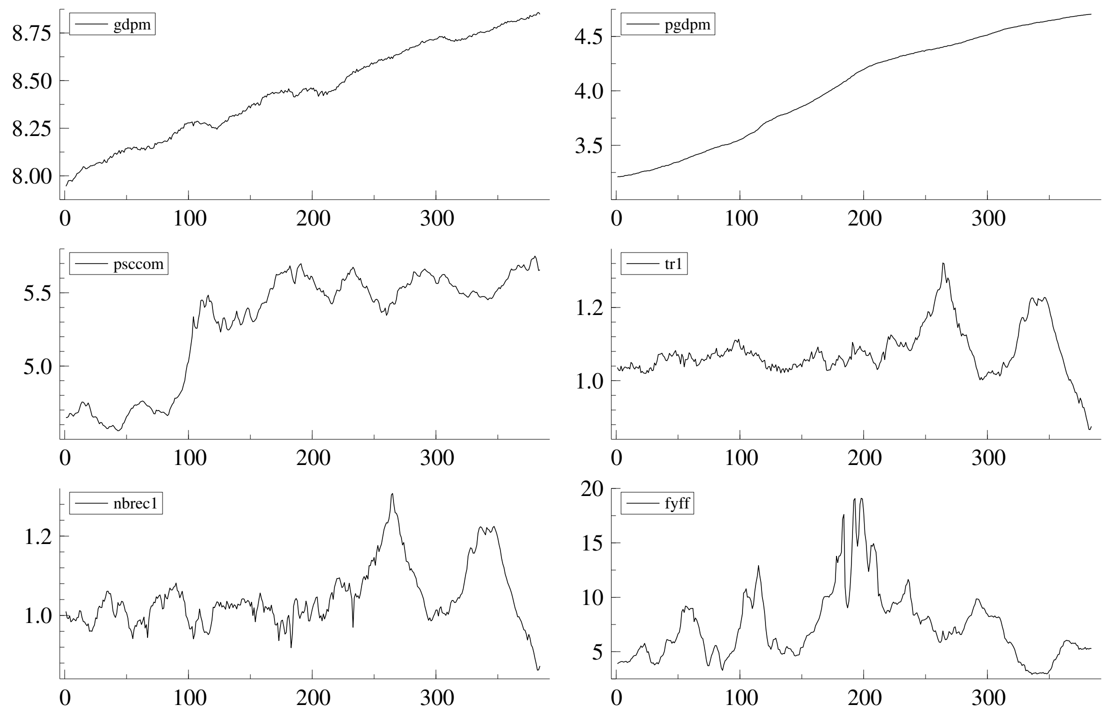


Figure 2. Impulse-response functions for VAR model.

A VAR(12) is fitted to the first difference of the six time series. The confidence band represent a one standard deviation. The standard deviations are derived from a parametric bootstrap using Gaussian innovations.

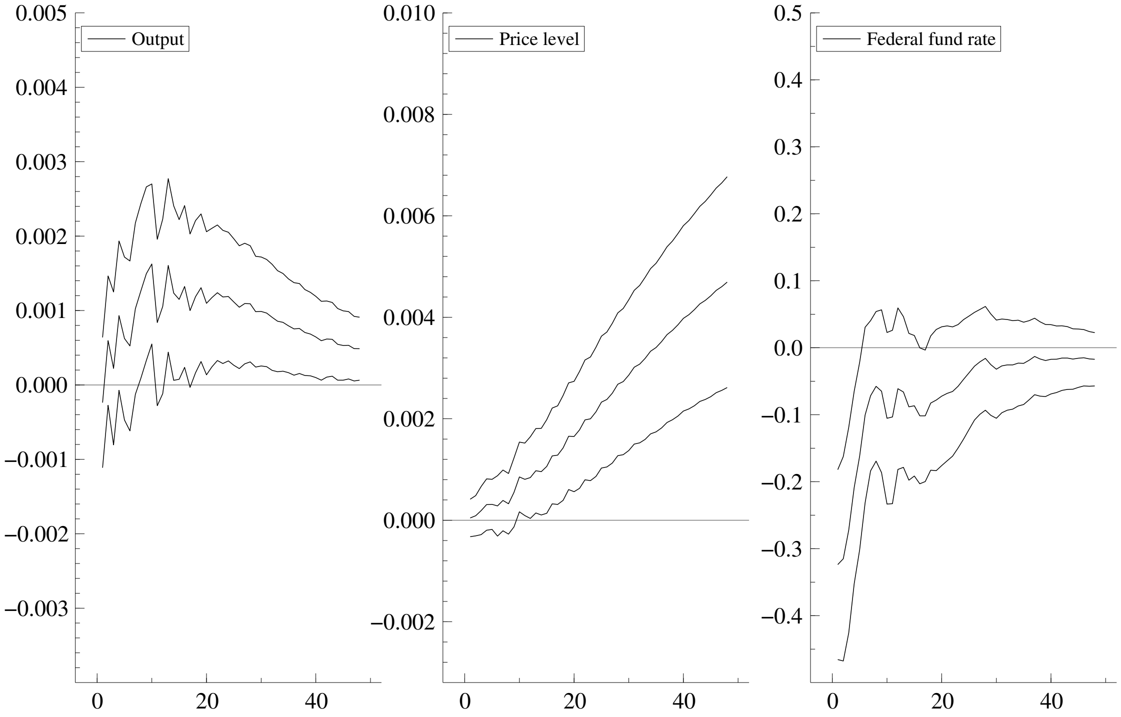
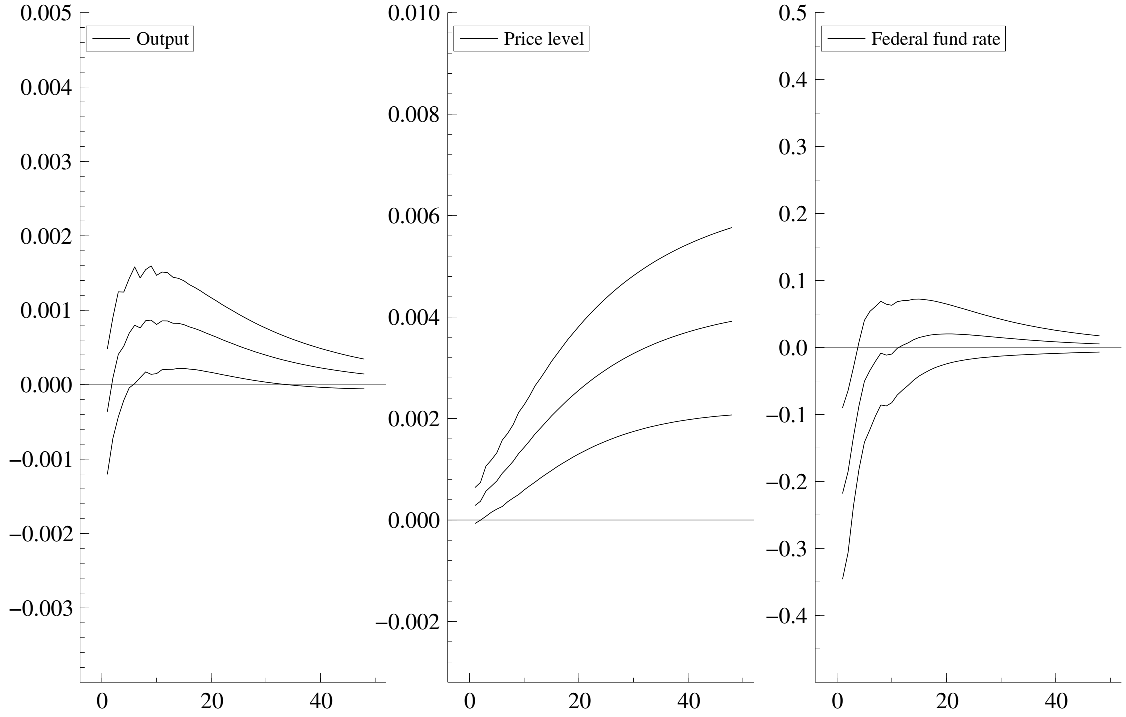


Figure 3. Impulse-response functions for VARMA model in final MA equation form.

A VARMA(3,10) is fitted to the first difference of the six time series. The confidence band represent a one standard deviation. The standard deviations are derived from a parametric bootstrap using Gaussian innovations.



of the regression-based estimation method proposed by Hannan and Rissanen (1982) for univariate ARMA models. Our method is in three steps. In a first step a long VAR is fitted to the data. In the second step, the lagged innovations in the VARMA model are replaced by the corresponding lagged residuals from the first step and a regression is performed. In a third step, the data from the second step are filtered and another regression is performed. We showed that the third-step estimators have the same asymptotic variance as their nonlinear counterpart (Gaussian maximum likelihood if the innovations are i.i.d., or generalized nonlinear least squares if they are merely uncorrelated). In the non i.i.d. case, we consider strong mixing conditions, rather than the usual martingale difference sequence assumption. We make these minimal assumptions on the innovations to broaden the class of models to which this method can be applied.

We also proposed a modified information criterion that gives consistent estimates of the orders of the AR and MA operators of the proposed VARMA representations. This criterion is to be minimized in the second step of the estimation method over a set of possible values for the different orders.

Monte Carlo simulation results indicates that the estimation method works well for small sample sizes and the information criterion picks the true value of the order p and q most of the time. These results holds for sample sizes commonly used in macroeconomics, *i.e.* 20 years of monthly data or 250 sample points. To demonstrate the importance of using VARMA models to study multivariate time series we compare the impulse-response functions and the out-of-sample forecasts generated by VARMA and VAR models when these models are applied to the dataset of macroeconomic time series used by Bernanke and Mihov (1998).

A. Proofs

Lemma A.1 *Let U and V be random variables measurable with respect to $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_{n}^{\infty}$, respectively where $\mathcal{F}_{a}^{b}$ is the σ-algebra of events generated by sets of the form $\{(X_{i_1}, X_{i_2}, \ldots, X_{i_n}) \in E_n\}$ with $a \leq i_1 < i_2 < \cdots < i_n \leq b$, and E_n is some n-dimensional Borel set. Let r_1, r_2, r_3 be positive numbers. Assume that $\|U\|_{r_1} < \infty$ and $\|V\|_{r_2} < \infty$ where $\|U\|_r = (\mathsf{E}[|U|]^r)^{1/r}$. If $r_1^{-1} + r_2^{-1} + r_3^{-1} = 1$, then there exists a positive constant c_0 independent of U, V and n, such that*

$$|\mathsf{E}[UV] - \mathsf{E}[U]\mathsf{E}[V]| \leq c_0 \|U\|_{r_1} \|V\|_{r_2} \alpha(n)^{1/r_3}.$$

where $\alpha(n)$ is defined in equation (2.15).

Proof. See Davydov (1968).

Lemma A.2 *If the random process (y_t) is strictly stationary and satisfies the strong mixing condition (2.15), with $\mathsf{E}|y_t|^{2+\delta} < \infty$ for some $\delta > 0$, and if $\sum_{j=1}^{\infty} \alpha(j)^{\delta/(2+\delta)} < \infty$, then*

$$\sigma^2 = \mathsf{E}\left[(y_t - \mathsf{E}[y_t])^2\right] + 2\sum_{j=1}^{\infty} \mathsf{E}\left[(y_t - \mathsf{E}[y_t])(y_{t+j} - \mathsf{E}[y_{t+j}])\right].$$

Moreover, if $\sigma \neq 0$ and $\mathsf{E}[y_t] = 0$, then

$$\Pr\left[\frac{y_1 + \cdots + y_t}{\sigma\sqrt{t}} < z\right] \underset{T\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du.$$

Proof. See Ibragimov (1962).

Proof of Lemma 3.8. Clearly, $\Phi(0) = \Theta(0) = I_K$ and $\det[\Phi(0)] = \det[\Theta(0)] = 1 \neq 0$. The polynomials $\det[\Phi(z)]$ and $\det[\Theta(z)]$ are different from zero in a neighborhood of zero. So we can choose $\rho_1 > 0$ such that $\det[\Phi(z)] \neq 0$ and $\det[\Theta(z)] \neq 0$ for $0 \leq |z| < \rho_1$. It follows that the matrices $\Phi(z)$ and $\Theta(z)$ are invertible for $0 \leq |z| < \bar{\rho}_0$, where $\bar{\rho}_0 = \min[\rho_0, \rho_1]$.

Let

$$C_0 = \{z \in \mathbb{C} \mid 0 \leq |z| < \bar{\rho}_0\}.$$

Then the matrices $\Phi(z)$ and $\Theta(z)$ are invertible for $z \in C_0$, and

$$\Phi(z)^{-1} = \frac{1}{\det[\Phi(z)]}\Phi^{\star}(z), \quad \Theta(z)^{-1} = \frac{1}{\det[\Theta(z)]}\Theta^{\star}(z),$$

where $\Phi^{\star}(z)$ and $\Theta^{\star}(z)$ are matrices of polynomials. Consequently, for $z \in C_0$, each element of $\Phi(z)^{-1}$ and $\Theta(z)^{-1}$ is a rational function whose denominator is different from zero. Thus, $\Phi(z)^{-1}$ and $\Theta(z)^{-1}$ are matrices of analytic functions on C_0, and the function

$$\Psi(z) = \Phi(z)^{-1}\Theta(z)$$

is analytic in the circle $0 \le |z| < \bar{\rho}_0$. Hence, it has a unique representation of the form

$$\Psi(z) = \sum_{k=0}^{\infty} \Psi_k z^k, \quad z \in C_0.$$

By assumption,

$$\Psi(z) = \Phi(z)^{-1}\Theta(z) = \bar{\Phi}(z)^{-1}\bar{\Theta}(z)$$

for $z \in C_0$. Hence, for $z \in C_0$,

$$\begin{aligned} \bar{\Phi}(z)\Phi(z)^{-1}\Theta(z) &= \bar{\Theta}(z), \\ \bar{\Phi}(z)\Phi(z)^{-1} &= \bar{\Theta}(z)\Theta(z)^{-1} \equiv \Delta(z), \end{aligned} \tag{A.1}$$

where $\Delta(z)$ is a diagonal matrix because $\Theta(z)$ and $\bar{\Theta}(z)$ are both diagonal,

$$\Delta(z) = \text{diag}\left[\delta_{ii}(z)\right],$$

where

$$\delta_{ii}(z) = \frac{\bar{\theta}_{ii}(z)}{\theta_{ii}(z)}, \ \theta_{ii}(0) = 1, \ \delta_{ii}(0) = \bar{\theta}_{ii}(0), \ i = 1, \ldots, K. \tag{A.2}$$

From (A.2), it follows that each $\delta_{ii}(z)$ is rational with no pole in C_0 such that $\delta_{ii}(0) = 1$, so it can be written in the form

$$\delta_{ii}(z) = \frac{e_i(z)}{f_i(z)}$$

where $e_i(z)$ and $f_i(z)$ have no common roots, $f_i(z) \neq 0$ for $z \in C_0$ and $\delta_{ii}(0) = e_{ii}(0) = 1$. From (A.1), it follows that for $z \in C_0$

$$\bar{\theta}_{ii}(z) = \delta_{ii}(z)\theta_{ii}(z), \quad \bar{\varphi}_{ij}(z) = \delta_{ii}(z)\varphi_{ij}(z), \quad i, j = 1, \ldots, K.$$

We first show that $\delta_{ii}(z)$ must be a polynomial. If $f_i(z) \neq 1$, then its order cannot be greater than the order $q_i \equiv \deg[\theta_{ii}(z)]$ for otherwise $\bar{\theta}_{ii}(z)$ would not be a polynomial. Similarly, if $f_i(z) \neq 1$ and is a polynomial of order less or equal to q_i, then all its roots must be roots of $\theta_{ii}(z)$ and $\varphi_{ij}(z)$, for otherwise $\bar{\theta}_{ii}(z)$ or $\bar{\varphi}_{ij}(z)$ would be a rational function. If $q_i \ge 1$, these roots are then common to $\theta_{ii}(z)$ and $\varphi_{ij}(z)$, $j = 1, \ldots, K$, which is in contradiction with Assumption **3.7**. Thus the degree of $f_i(z)$ must be zero, and $\delta_{ii}(z)$ is a polynomial.

If $\delta_{ii}(z)$ is a polynomial of degree greater than zero, this would entail that $\bar{\theta}_{ii}(z)$ and $\bar{\varphi}_{ij}(z)$ have roots in common, in contradiction with Assumption **3.7**. Thus $\delta_{ii}(z)$ must be a constant. Further, $\delta_{ii}(0) = 1$ so that for $i = 1, \ldots, K$,

$$\bar{\theta}_{ii}(z) = \theta_{ii}(z), \quad \bar{\varphi}_{ij}(z) = \varphi_{ij}(z), \quad j = 1 \ldots, K,$$

hence

$$\bar{\Theta}(z) = \Theta(z), \quad \bar{\Phi}(z) = \Phi(z).$$

Proof of Theorem 3.9. Under the assumption that the VARMA process is invertible, we can write

$$\Theta(L)^{-1}\Phi(L)Y_t = U_t.$$

Now suppose by contradiction that there exist operators $\bar{\Phi}(L)$ and $\bar{\Theta}(L)$, with $\bar{\Theta}(L)$ diagonal and invertible, and $\bar{\Phi}(L) \neq \Phi(L)$ or $\bar{\Theta}(L) \neq \Theta(L)$, such that

$$\bar{\Theta}(L)^{-1}\bar{\Phi}(L) = \Theta(L)^{-1}\Phi(L),$$

If the above equality hold, then it must also be the case that

$$\bar{\Theta}(z)^{-1}\bar{\Phi}(z) = \Theta(z)^{-1}\Phi(z), \quad \forall z \in C_0,$$

where $C_0 = \{z \in \mathbb{C} \mid 0 \leq |z| < \rho_0\}$ and $\rho_0 > 0$. By Lemma **3.8**, it follows that

$$\bar{\Phi}(z) = \Phi(z), \quad \bar{\Theta}(z) = \Theta(z) \quad \forall z.$$

Hence, the representation is unique.

Proof of Theorem 3.11. The proof can be easily adapted from the proof of Theorem **3.9** once we replace Assumption **3.7** by Assumption **3.10**.

Lemma A.3 (Infinite VAR convergence) *If the VARMA model is invertible and if n_T grows at a rate faster then $\log T$, then $\sum_{j=1}^{K}\sum_{m=n_T+1}^{\infty}|\pi_{ij,m}| = o(T^{-1})$ for $i = 1, \ldots, K$.*

Proof of Lemma A.3. The matrix $\Theta(L)^{-1}$ can be seen has its adjoint matrix divided by its determinant. Since Y_t is invertible, the roots of $\det\Theta(L)$ are outside the unit circle and so the elements of $\Pi(L) = \Theta(L)^{-1}\Phi(L)$ decrease exponentially:

$$|\pi_{ij,m}| \leq c\rho^m, \ \forall i, j.$$

with $c > 0$ and $0 < \rho < 1$. From this

$$\begin{aligned} \sum_{j=1}^{K}\sum_{l=n_T+1}^{T}|\pi_{ij,m}| &\leq \sum_{j=1}^{K}\sum_{l=n_T+1}^{T}c\rho^m \\ &= c\,K\,\frac{\rho^{n_T+1}}{1-\rho}. \end{aligned}$$

If n_T grows at a rate faster than $\log T$ then $T\rho^{n_T}$ will tend to zero since $|\rho| < 1$.

Lemma A.4 (Covariance estimation) *If the process $\{Y_t\}$ is a strictly stationary VARMA process with $\{U_t\}$ uncorrelated, $\mathsf{E}[|u_{it}|^{4+2\delta}] < \infty$ for some $\delta > 0$, α-mixing with $\sum_{h=1}^{\infty}\alpha(h)^{2/(2+\delta)} < \infty$*

then

$$\frac{1}{T}\sum_{t=1}^{T} y_{k,t-r}y_{k',t-s} - \mathsf{E}[y_{k,t-r}y_{k',t-s}] = O_p(T^{-1/2}) \ \ \forall k, k'.$$

Proof of Lemma A.4. First notice that by stationarity,

$$\mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T} y_{k,t-r}y_{k',t-s}\right] - \mathsf{E}[y_{k,t-r}y_{k',t-s}] = 0.$$

Now taking the variance,

$$\begin{aligned}
& Var\left[\frac{1}{T}\sum_{t=1}^{T} y_{t-r}(k)y_{t-s}(k')\right] \\
= \quad & \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} Cov\left[y_{k,t-r}y_{k',t-s}\,;\,y_{k,t'-r}y_{k',t'-s}\right] \\
= \quad & \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=t+s-r+1}^{T} Cov\left[y_{k,t-r}y_{k',t-s}\,;\,y_{k,t'-r}y_{k',t'-s}\right] \\
+ \quad & \frac{1}{T^2}\sum_{t'=1}^{T}\sum_{t=t'+s-r+1}^{T} Cov\left[y_{k,t-r}y_{k',t-s}\,;\,y_{k,t'-r}y_{k',t'-s}\right] \\
+ \quad & \frac{1}{T^2}\sum_{t'=1+(s-r)}^{T-(s-r)}\sum_{t=t'-(s-r)}^{t'+(s-r)} Cov\left[y_{k,t-r}y_{k',t-s}\,;\,y_{k,t'-r}y_{k',t'-s}\right] \\
+ \quad & \frac{1}{T^2}\sum_{t'=1}^{(s-r+1)}\sum_{t=1}^{t'+(s-r)} Cov\left[y_{k,t-r}y_{k',t-s}\,;\,y_{k,t'-r}y_{k',t'-s}\right] \\
+ \quad & \frac{1}{T^2}\sum_{t=T-(s-r)}^{T}\sum_{t'=t-(s-r)}^{t+(s-r)} Cov\left[y_{k,t-r}y_{k',t-s}\,;\,y_{k,t'-r}y_{k',t'-s}\right].
\end{aligned} \tag{A.3}$$

For the first two terms of equation (A.3), Using Davydov's inequality (lemma **A.1**), the strong mixing hypothesis and the finite fourth moment we know that

$$\begin{aligned}
& \sum_{t'=t+s-r+1}^{T} \left|Cov\left[y_{k,t-r}y_{k',t-s}\,;\,y_{k,t'-r}y_{k',t'-s}\right]\right| \\
\leq \quad & \sum_{t'=t+s-r+1}^{T} \left(c_0 \,\|(y_{k,t-r}y_{k',t-s}\|_{2+\delta}\|y_{k,t'-r}y_{k',t'-s}\|_{2+\delta}\,\alpha(t'-t-s+r-1)^{\delta/(2+\delta)}\right) \\
< \quad & \infty
\end{aligned}$$

from which we conclude that the first two terms converge to zero at rate $1/T$. For the other three terms, since these covariances are finite, the sums divided by T^2 will also converge to zero at rate $1/T$. Hence,

$$\frac{1}{T}\sum_{t=1}^{T} y_{k,t-r}y_{k',t-s} - \mathsf{E}[y_{k,t-r}y_{k',t-s}] = O_p(T^{-1/2}) \ \ \forall k, k'.$$

Proof of Theorem 4.1. We first introduce some additional matrix norms:

$$\begin{aligned}
\|B\|_2^2 &= \sup_{l\neq 0}\frac{l'B'Bl}{l'l}, &\text{(A.4)}\\
\|B\|_1 &= \max_{i\leq j\leq n}\sum_{i=1}^{n}|b_{ij}|, &\text{(A.5)}\\
\|B\|_\infty &= \max_{1\leq i\leq n}\sum_{j=1}^{n}|b_{ij}|, &\text{(A.6)}
\end{aligned}$$

where (A.4) is the largest eigenvalue of $B'B$. Useful inequalities relating these norms are given in Horn and Johnson (1985, p. 313):

$$\|AB\|^2 \leq \|A\|_2^2\|B\|^2\ , \quad \|AB\|^2 \leq \|A\|^2\|B\|_2^2\ , \quad \|B\|_2^2 \leq \|B\|_1\|B\|_\infty. \tag{A.7}$$

In the first step estimation, we regress

$$y_{it} = \sum_{l=1}^{n_T}\sum_{j=1}^{K}\pi_{ij,l}y_{j,t-l} + e_{it} \tag{A.8}$$

when in fact

$$y_{it} = \sum_{l=1}^{\infty}\sum_{j=1}^{K}\pi_{ij,l}y_{j,t-l} + u_{it}.$$

If we let

$$\hat{B}(n_T) = \sum_{t=n_T+1}^{T}\frac{\mathbf{Y}_{t-1}'(n_T)\mathbf{Y}_{t-1}(n_T)}{T-n_T},$$

then OLS applied to (A.8) yields:

$$\begin{aligned}
\hat{\Pi}_{i\bullet}(n_T) &= [\hat{\pi}_{i\bullet,1},\ \ldots\ ,\hat{\pi}_{i\bullet,n_T}]'\\
&= \hat{B}(n_T)^{-1}\sum_{t=n_T+1}^{T}\frac{\mathbf{Y}_{t-1}(n_T)'y_{it}}{T-n_T}\\
&= \hat{B}(n_T)^{-1}\sum_{t=n_T+1}^{T}\frac{\mathbf{Y}_{t-1}(n_T)'}{T-n_T}\left\{\sum_{l=1}^{\infty}\pi_{i\bullet,l}Y_{t-l} + u_{it}\right\}
\end{aligned}$$

$$= \quad \mathit{\Pi}_{i\bullet}(n_T) + \hat{B}(n_T)^{-1} \sum_{t=n_T+1}^{T} \frac{\mathbf{Y}_{t-1}(n_T)'}{T-n_T} \left\{ \sum_{l=n_T+1}^{\infty} \pi_{i\bullet,l} Y_{t-l} + u_{it} \right\}.$$

Rearranging the elements,

$$\begin{aligned}
\hat{\mathit{\Pi}}_{i\bullet}(n_T) - \mathit{\Pi}_{i\bullet}(n_T) &= \hat{B}(n_T)^{-1} \sum_{t=n_T+1}^{T} \frac{\mathbf{Y}'_{t-1}(n_T)}{T-n_T} \left\{ \sum_{l=n_T+1}^{\infty} \pi_{i\bullet,l} Y_{t-l} \right\} + \\
&\quad \hat{B}(n_T)^{-1} \sum_{t=n_T+1}^{T} \frac{\mathbf{Y}'_{t-1}(n_T) u_{it}}{T-n_T}
\end{aligned}$$

Using inequalities (A.7) and the fact that $\hat{B}(n_T)$ is symmetric,

$$\begin{aligned}
\|\hat{\mathit{\Pi}}_{i\bullet}(n_T) - \mathit{\Pi}_{i\bullet}(n_T)\| &\leq \|\hat{B}(n_T)^{-1}\|_2 \|V_{1T}\| + \|\hat{B}(n_T)^{-1}\|_2 \|V_{2T}\| \\
&\leq \|\hat{B}(n_T)^{-1}\|_1 \|V_{1T}\| + \|\hat{B}(n_T)^{-1}\|_1 \|V_{2T}\| \qquad \text{(A.9)}
\end{aligned}$$

where

$$\begin{aligned}
V_{1T} &= \frac{1}{T-n_T} \sum_{t=n_T+1}^{T} \mathbf{Y}'_{t-1}(n_T) \sum_{l=n_T+1}^{\infty} \pi_{i\bullet,l} Y_{t-l}, \\
V_{2T} &= \frac{1}{T-n_T} \sum_{t=n_T+1}^{T} \mathbf{Y}'_{t-1}(n_T) u_{it}.
\end{aligned}$$

Firstly, $\|V_{2T}\|^2$ can be expanded into

$$\begin{aligned}
\|V_{2T}\|^2 &= tr\left(V'_{2T} V_{2T}\right) \\
&= \sum_{k=1}^{K} \sum_{l=1}^{n_T} \left(\frac{\sum_{t=n_T+1}^{T} y_{k,t-l} u_{it}}{T-n_T} \right)^2 \\
&= \sum_{k=1}^{K} \sum_{l=1}^{n_T} \left(\underbrace{\mathsf{E}[y_{k,t-l} u_{it}]}_{=0} + O_p(T^{-1/2}) \right)^2.
\end{aligned}$$

It follows that $\|V_{2T}\| = O_p(n_T T^{-1/2})$. Similarly, for $\|V_{1T}\|^2$

$$\begin{aligned}
\|V_{1T}\|^2 &= tr\left(V'_{1T} V_{1T}\right) \\
&= \sum_{k=1}^{K} \sum_{l=1}^{n_T} \left(\frac{\sum_{t=n_T+1}^{T} y_{k,t-l} [\sum_{m=n_T+1}^{\infty} \sum_{j=1}^{K} \pi_{ij,m} y_{j,t-m}]}{T-n_T} \right)^2
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{K}\sum_{l=1}^{n_T}\left(\sum_{j=1}^{K}\sum_{m=n_T+1}^{\infty}\pi_{ij,m}\frac{1}{T-n_T}\sum_{t=n_T+1}^{T}y_{k,t-l}y_{j,t-m}\right)^2 \\
&= \sum_{k=1}^{K}\sum_{l=1}^{n_T}\left(\sum_{j=1}^{K}\sum_{m=n_T+1}^{\infty}\pi_{ij,m}[Cov(y_{k,t-l};y_{j,t-m})+O_p(T^{-1/2})]\right)^2 .
\end{aligned}
$$

Since Y_t is invertible, $|\pi_{ij,m}| < c_0\rho^m$ and $\sum_{m=n_T+1}^{\infty}|\pi_{ij,m}| < c_0\rho^{n_T}/(1-\rho)$. It follows that $\sum_{m=n_T+1}^{\infty}\pi_{ij,m}[Cov(y_{k,t-l};y_{j,t-m})+O_P(T^{-1/2})] = O_p(T^{-1/2})$ since $n_T/\log(T)\to 0$ as $T\to\infty$. Hence, $\|V_{1T}\| = O_p(n_T T^{-1/2})$.

For $\|\hat{B}(n_T)^{-1}\|_1$, the existence of $\hat{B}(n_T)^{-1}$ is guaranteed by a lemma that can be found in Tiao and Tsay (1983). The argument is the following. It is clear that $\hat{B}(n_T)$ is a symmetric non-negative definite matrix. To show that it is positive definite take $c=[c_1,\dots c_{k\,n_T}]'$ be any arbitrary vector and consider

$$
c'\hat{B}(n_T)c = \frac{1}{T^2}\sum_{t=n_T+1}^{T}\left(\sum_{j=1}^{n_T}\sum_{k=1}^{K}c_{(j-1)K+k}y_{k,t-j}\right)^2 .
$$

If $c'\hat{B}(n_T)c=0$, then

$$
\sum_{j=1}^{n_T}\sum_{k=1}^{K}c_{(j-1)K+k}y_{k,t-j}=0 \quad \text{for} \quad t=n_T+1,\,\dots\,,T,
$$

which since $T>2:K:n_T$, is a system of linear equations of $K:n_T$ unknowns and at least $K:n_T$ equations. Since Y_t is continuous and non deterministic, this implies that $c=0$. This proves that $\hat{B}(n_T)$ is positive definite. Denoting by $B(n_T)$ the $(K:n_T\times K:n_T)$ matrix of the corresponding covariances instead of the empirical covariances, we can use a similar argument to show that $B(n_T)$ is also positive definite, where

$$
B(n_T) = \begin{bmatrix} \varGamma(0) & \varGamma(1) & \cdots & \varGamma(n_T-1) \\ \varGamma(-1) & \varGamma(0) & \cdots & \varGamma(n_T-2) \\ \vdots & & \ddots & \vdots \\ \varGamma(-n_T+1) & \varGamma(-n_T+2) & \cdots & \varGamma(0) \end{bmatrix}
$$

and $\varGamma(h)=Cov(Y_t;Y_{t-h})$.

We also know that since Y_t is stationary, *i.e.* the roots of $\det(P(L))$ are outside the unit circle, $\sum_{h=0}^{\infty}|Cov(y_{i,t};y_{j,t-h})|<\infty$. It follows that the sum of the elements along a row of $B(n_T)$ is uniformly bounded in n_T and the row number. This property must also holds for $B(n_T)^{-1}$. If it was not the case we would have

$$
\begin{aligned}
B(n_T)^{-1}B(n_T) &= I_{n_T} \\
B(n_T)^{-1}B(n_T)\mathbf{i}_{n_T} &= \mathbf{i}_{n_T}
\end{aligned}
$$

with $\mathbf{i}_{n_T}$ an $(K\ n_T \times 1)$ vector of ones. Since $B(n_T)\mathbf{i}_{n_T}$ gives a vector of bounded elements, $B(n_T)^{-1}$ must have the sum of elements bounded along any row.

From lemma **A.4** we know that each element of $\hat{B}(n_T) - B(n_T)$ is $O_P(T^{-1/2})$. It follows that

$$\sum_{h=1}^{Kn_T} \hat{B}_{ih}(n_T) = \sum_{h=1}^{Kn_T} B_{ih}(n_T) + O_p(n_T T^{-1/2}).$$

Using a similar argument as above, the sum of the elements along a row of $\hat{B}_{n_T}$ must also be uniformly bounded in n_T and the row number and $\|\hat{B}(n_T)^{-1}\|_1 < c_0 + O_p(n_T T^{-1/2})$. Hence, $\|\hat{\Pi}_{i\bullet}(n_T) - \Pi_{i\bullet}(n_T)\| = O_p(n_T T^{-1/2})$.

Proof of Theorem 4.2. If we denote by Z_{t-1} the equivalent of $\hat{Z}_{t-1}$ which contains the true innovations u_{kt} instead of the residuals $\hat{u}_{kt}$,

$$\begin{aligned}
\hat{\gamma} &= \left[\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}_U^{-1}\hat{Z}_{t-1}\right]^{-1} \left[\sum_{t=l}^{T} \hat{Z}_{t-1}\hat{\Sigma}_U^{-1}\left(Z_{t-1}\gamma + U_t\right)\right] \\
&= \left[\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}_U^{-1}\hat{Z}_{t-1}\right]^{-1} \left[\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}_U^{-1}Z_{t-1}\right]\gamma + \\
&\quad \left[\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}_U^{-1}\hat{Z}_{t-1}\right]^{-1} \left[\sum_{t=l}^{T} \hat{Z}_{t-1}\hat{\Sigma}_U^{-1}U_t\right].
\end{aligned}$$

Firstly, we show that $\hat{\Sigma}_U \xrightarrow{p} \Sigma_U$. We can write the residual $\hat{U}_t$ as

$$\begin{aligned}
\hat{U}_t &= \hat{\Pi}^{n_T}(L)Y_t \\
&= \hat{\Pi}^{n_T}(L)\Psi(L)U_t \\
&= [I_K + (\hat{\Pi}^{n_T}(L)\Psi(L) - I_K)]U_t \\
&= [I_K + (\hat{\Pi}^{n_T}(L) - \Pi(L))\Psi(L)]U_t \\
&= U_t + (\hat{\Pi}^{n_T}(L) - \Pi(L))Y_t\,.
\end{aligned}$$

Using the results from Theorem **4.1** where we showed that $\sum_{l=1}^{n_T} \|\hat{\Pi}_l^{n_T} - \Pi_l\| = O_p(n_T T^{-1/2})$ combined with $\sum_{l=n_T+1}^{\infty} \|\Pi_l\| = o(T^{-1})$ if $\log(T)/n_T \to 0$, we can conclude that

$$\begin{aligned}
\hat{\Sigma}_U = \frac{1}{T-n_T}\sum_{t=n_T+1}^{T} \hat{U}_t\hat{U}'_t &= \frac{1}{T-n_T}\sum_{t=n_T+1}^{T} U_tU'_t + o_p(T^{-1/2}) \\
&\xrightarrow{p} \Sigma_U\,.
\end{aligned}$$

To show that $\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}_U^{-1}\hat{Z}_{t-1}/T$ converge to $\tilde{J} = \mathsf{E}[Z'_{t-1}\Sigma_U^{-1}Z_{t-1}]$ in probability, since $\hat{\Sigma}_U \xrightarrow{p} \Sigma_U$ we only have to show that

- $\sum_{t=l}^{T} y_{i,t-k}y_{j,t-l}/T \xrightarrow{p} \mathsf{E}[y_{i,t-k}y_{j,t-l}]$,
- $\sum_{t=l}^{T} \hat{u}_{i,t-k}\hat{u}_{j,t-l}/T \xrightarrow{p} \mathsf{E}[u_{i,t-k}u_{j,t-l}]$,
- $\sum_{t=l}^{T} y_{i,t-k}\hat{u}_{j,t-l} \xrightarrow{p} \mathsf{E}[y_{i,t-k}u_{j,t-l}]$.

The first is proved in lemma **A.4** and the second can be proved in a similar manner. We can easily prove the third by using results for the previous two and Theorem **4.1**. Similarly, $\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}_U^{-1}Z_{t-1}/T$ converge also to $\tilde{J}$. Using similar calculus, we can also show that $\sum_{t=l}^{T} \hat{Z}_{t-1}\hat{\Sigma}_U^{-1}U_t/T = O_p(T^{-1/2})$. Combining all these results we can conclude that $\tilde{\gamma} \xrightarrow{p} \gamma$.

For the asymptotic distribution, using Ibragimov's central limit theorem we can conclude that

$$\frac{1}{\sqrt{T}}\sum_{t=l}^{T} \hat{Z}'_{t-1}\hat{\Sigma}_U^{-1}U_t \xrightarrow{d} \mathcal{N}\left(0, \hat{I}\right)$$

with

$$\tilde{I} = \sum_{j=-\infty}^{\infty} \mathsf{E}\left[\left\{Z'_{t-1}\Sigma_U^{-1}U_t\right\}\left\{Z'_{t-1-j}\Sigma_U^{-1}U_{t-j}\right\}\right].$$

From this,

$$\sqrt{T}\left(\tilde{\gamma} - \gamma\right) \xrightarrow{d} \mathcal{N}\left(0, \tilde{J}^{-1}\tilde{I}\tilde{J}^{-1}\right).$$

From the preceding results, it is obvious that $\tilde{J}$ can be consistently estimated by

$$\tilde{J}_T = \frac{1}{T}\sum_{t=n_T+1}^{T} \hat{Z}'_{t-1}\hat{\Sigma}_U^{-1}\hat{Z}_{t-1}$$

and using theorem 2 of Newey and West (1987), we know that $\tilde{I}_T \xrightarrow{p} \tilde{I}$ if we take $m_T^4/T \to 0$ with $m_T \to \infty$ as $T \to \infty$.

Proof of Theorem 4.3. First we can rewrite X_t, W_t and $\tilde{V}_t$ as

$$X_t = \hat{\Theta}(L)^{-1}Y_t\,, \quad W_t = \hat{\Theta}(L)^{-1}\tilde{U}_t\,, \quad \tilde{V}_t = \hat{\Theta}(L)^{-1}\tilde{Z}_t.$$

We can also rewrite $\tilde{U}_t + X_t - W_t$ as

$$\begin{aligned}
\tilde{U}_t + X_t - W_t &= \hat{\Theta}(L)^{-1}Y_t + \tilde{U}_t - \hat{\Theta}(L)^{-1}\tilde{U}_t \\
&= \hat{\Theta}(L)^{-1}\left[Z_{t-1}\gamma + U_t\right] + \tilde{U}_t - \hat{\Theta}(L)^{-1}\tilde{U}_t \\
&= \hat{\Theta}(L)^{-1}Z_{t-1}\gamma + \hat{\Theta}(L)^{-1}U_t + \tilde{U}_t - \hat{\Theta}(L)^{-1}\tilde{U}_t \\
&= V_{t-1}\gamma + U_t + [(\tilde{U}_t - U_t) - \hat{\Theta}(L)^{-1}(\tilde{U}_t - U_t)] \\
&= V_{t-1}\gamma + U_t + O_p(T^{-1/2}).
\end{aligned}$$

With this, the regression becomes

$$\begin{aligned}
\hat{\gamma} &= \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}' \tilde{\Sigma}_U^{-1} \tilde{V}_{t-1}\right]^{-1} \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}' \tilde{\Sigma}_U^{-1} \left(\tilde{U}_t + X_t - W_t\right)\right] \\
&= \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}' \tilde{\Sigma}_U^{-1} \tilde{V}_{t-1}\right]^{-1} \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}' \tilde{\Sigma}_U^{-1} V_{t-1}\right] \gamma + \\
&\quad \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}' \tilde{\Sigma}_U^{-1} \tilde{V}_{t-1}\right]^{-1} \left[\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}' \tilde{\Sigma}_U^{-1} U_t\right] + o_p(T^{-1/2}).
\end{aligned}$$

Just like in the proof of theorem **4.2** we see that $\hat{\gamma} - \gamma = O_p(T^{-1/2})$. Using Ibragimov's central limit theorem we conclude that

$$\sqrt{T}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}\left(0, \hat{J}^{-1}\hat{I}\hat{J}\right)$$

with

$$\begin{aligned}
\hat{I} &= \sum_{j=-\infty}^{\infty} \mathsf{E}\left[\left\{V_{t-1}\Sigma_U^{-1}U_t\right\}\left\{V_{t-1-j}\Sigma_U^{-1}U_{t-j}\right\}\right], \\
\hat{J} &= \mathsf{E}\left[V_{t-1}'\Sigma_U^{-1}V_{t-1}\right]
\end{aligned}$$

and just like in the proof of theorem **4.2** the matrix $\hat{I}$ and $\hat{J}$ can be consistently estimated respectively by

$$\begin{aligned}
\hat{I}_T &= \frac{1}{T}\sum_{j=-m_T}^{m_T} \omega(j, m_T) \sum_{t=\max(p,q)+1+|j|}^{T} \left\{\tilde{V}_{t-1}\tilde{\Sigma}_U^{-1}\tilde{U}_t\right\}\left\{\tilde{V}_{t-1-j}\tilde{\Sigma}_U^{-1}\tilde{U}_{t-j}\right\}, \\
\hat{J}_T &= \frac{1}{T}\sum_{t=\max(p,q)+1}^{T} \tilde{V}_{t-1}'\tilde{\Sigma}_U^{-1}\tilde{V}_{t-1}.
\end{aligned}$$

Proof of Theorem 5.1. Take the difference between the information criterion for given values of the orders p and q, and its true value (for the true values p_0, q_0 and Σ_0)

$$\log(\det \tilde{\Sigma}_U) - \log(\det \Sigma_0) + [\dim \gamma(p,q) - \dim \gamma(p_0, q_0)] \frac{(\log T)^{1+\delta}}{T}.$$

First, consider the case where $p < p_0$ or $q < q_0$. In this case, as T grows to infinity, eventually $\log(\det \tilde{\Sigma}_U) > \log(\det \Sigma_0)$ because of the left-coprime property. So eventually we must have $p \geq p_0$ and $q \geq q_0$ because the difference of the two criteria will be positive.

Before studying the case $p \geq p_0$ or $q \geq q_0$, let us write the residuals $\tilde{U}_t$ for general lag operators $\Phi(L)$ and $\Theta(L)$ in the following way:

$$\begin{aligned}
\tilde{U}_t &= \Phi(L)Y_t - (\Theta(L) - I_K)\hat{U}_t \\
&= \Phi(L)Y_t - (\Theta(L) - I_K)\hat{\Pi}^{n_T}(L)Y_t \\
&= \left[\Phi(L) - (\Theta(L) - I_K)\hat{\Pi}^{n_T}(L)\right] Y_t \\
&= \left[\Phi(L) - (\Theta(L) - I_K)\Theta_0(L)^{-1}\Theta_0(L)\hat{\Pi}^{n_T}(L)\right] Y_t \\
&= \left[\chi(L) - (\Theta(L) - I_K)\Theta_0(L)^{-1}C(L) + \Theta_0(L)^{-1}\Phi_0(L)\right] Y_t
\end{aligned} \tag{A.10}$$

where $C(L) = \Theta_0(L)\hat{\Pi}^{n_T}(L) - \Phi_0(L)$ and $\chi(L) = \Phi(L) - \Theta(L)\Theta_0(L)^{-1}\Phi_0(L)$.

For the case $p = p_0$ and $q = q_0$, using (A.10), it follows that

$$\begin{aligned}
\tilde{\Sigma}_U &= \inf_{\Phi,\Theta} \frac{1}{T} \sum_{t=n_T+1}^{T} \tilde{U}_t \tilde{U}_t' \\
&= \inf_{\Phi,\Theta} \frac{1}{T} \sum_{t=n_T+1}^{T} \left[\chi(L) - (\Theta(L) - I_K)\Theta_0(L)^{-1}C(L) + \Theta_0(L)^{-1}\Phi_0(L)\right] Y_t Y_t' \times \\
&\qquad \left[\chi(L) - (\Theta(L) - I_K)\Theta_0(L)^{-1}C(L) + \Theta_0(L)^{-1}\Phi_0(L)\right]'.
\end{aligned}$$

From previous calculus we know that

$$\begin{aligned}
& \frac{1}{T} \sum_{t=n_T+1}^{T} \left\{\Theta_0(L)^{-1}\Phi_0(L)Y_t\right\} \left\{\Theta_0(L)^{-1}\Phi_0(L)Y_t\right\}' \\
=\; & \frac{1}{T} \sum_{t=n_T+1}^{T} U_t U_t' \\
=\; & \Sigma_U + O_p(T^{-1/2}).
\end{aligned}$$

For the cross-product involving $\chi(L)$ we see that

$$\begin{aligned}
& \inf_{\Phi,\Theta} \frac{1}{T} \sum_{t=n_T+1}^{T} \{\chi(L)Y_t\} \{\chi(L)Y_t\}' \\
=\; & \inf_{\Phi,\Theta} \frac{1}{T} \sum_{t=n_T+1}^{T} \left\{[\Phi(L) - \Theta(L)\Theta_0(L)^{-1}\Phi_0(L)]Y_t\right\} \left\{[\Phi(L) - \Theta(L)\Theta_0(L)^{-1}\Phi_0(L)]Y_t\right\}' \\
=\; & \inf_{\Phi,\Theta} \frac{1}{T} \sum_{t=n_T+1}^{T} \left\{[\Phi(L)\Phi_0(L)^{-1}\Theta_0(L) - \Theta(L)]U_t\right\} \left\{[\Phi(L)\Phi_0(L)^{-1}\Theta_0(L) - \Theta(L)]U_t\right\}' \\
=\; & O_p(T^{-1})
\end{aligned}$$

because $\Phi(L)\Phi_0(L)^{-1}\Theta_0(L) - \Theta(L)$ is $O_p(T^{-1/2})$. We have similar results for all the remaining cross-product except for

$$\frac{1}{T}\sum_{t=n_T+1}^{T}\left\{(\Theta(L)-I_K)\Theta_0(L)^{-1}C(L)Y_t\right\}\left\{(\Theta(L)-I_K)\Theta_0(L)^{-1}C(L)Y_t\right\}'.$$

We saw previously that $\sum_{l=1}^{n_T}\|\hat{\Pi}_l(n_T)-\Pi_l\| = O_p(n_T T^{-1/2})$ so we have the same result for $C(L)$, *i.e.* $\sum_{l=1}^{n_T+q}\sum_{j=1}^{K}|C_{ij,l}| = O_p(n_T T^{-1/2})$. Combining this with the fact that $\frac{1}{T}\sum_{t=n_T+1}^{T} Y_t Y_t' - \mathsf{E}[Y_t Y_t'] = O_p(T^{-1/2})$ we can conclude that

$$\frac{1}{T}\sum_{t=n_T+1}^{T}\left\{[\Theta(L)-I_K]\Theta_0(L)^{-1}C(L)Y_t\right\}\left\{[\Theta(L)-I_K]\Theta_0(L)^{-1}C(L)Y_t\right\}' = O_p(n_T T^{-1}).$$

Combining these results we see that for $p = p_0$ and $q = q_0$, $\tilde{\Sigma} = \Sigma_0 + O_p(n_T T^{-1})$ and equivalently

$$\det\tilde{\Sigma} = \det\Sigma_0 + O_p(n_T T^{-1}).$$

For the case where $p \geq p_0$, $q \geq q_0$ with either p or q greater then their true value, even though the model might not be identified in this case, for the minimization of $\det\tilde{\Sigma}$ we can not do any worse than in the case where $p = p_0$, $q = q_0$ so the infimum will yield the same result than for the case $p = p_0$ and $q = q_0$. So eventually

$$\det\tilde{\Sigma}_U - \det\Sigma_0 + [\dim\gamma(p,q) - \dim\gamma(p_0,q_0)]\,\frac{(\log T)^{1+\delta}}{T} \geq 0$$

because the penalty on the number of parameters will dominate. So if to select the order p and q we use an information criterion such as

$$\log(\det\hat{\Sigma}_U) + (\dim\gamma)\frac{(\log T)^{1+\delta}}{T}$$

with $\delta > 0$, we will get $\hat{p} \to p_0$, $\hat{q} \to q_0$ since $\log(\det\hat{\Sigma}_U) - \log(\det\Sigma_0) = O_p(n_T T^{-1})$.

Proof of Theorem 5.2. The proof is similar to the proof of Theorem **5.1** since we easily see that $\hat{\sigma}^2 = \sigma_0^2 + O_p(n_T T^{-1})$.

References

BARTEL, H., AND LÜTKEPOHL (1998): "Estimating the Kronecker indices of cointegrated echelon-form VARMA models," *Econometrics Journal*, 1, C76–C99.

BERNANKE, B. S., AND I. MIHOV (1998): "Measuring Monetary Policy," *The Quarterly Journal of Economics*, 113(3), 869–902.

BOSQ, D. (1998): *Nonparametric Statistics for Stochastic Processes - Estimation and Prediction*, no. 110 in Lecture Notes in statistics. Springer-Verlag, Berlin, second edn.

BOX, G. E. P., AND D. A. PIERCE (1970): "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," *JASA*, 65, 1509–1526.

BROCKWELL, P. J., AND R. A. DAVIS (1991): *Time series: theory and methods*, Springer Series in Statistics. Springer-Verlag, New York, second edn.

COOPER, D. M., AND E. F. WOOD (1982): "Identifying Multivariate Time Series Models," *Journal of Time Series Analysis*, 3, 153–164.

DAVYDOV, Y. A. (1968): "Convergence of Distributions Generated by Stationary Stochastic Processes," *Theory of Probability and its Applications*, pp. 691–696.

DEISTLER, M., AND E. J. HANNAN (1981): "Some Properties of the Parametrization of ARMA Systems with Unknown Order," *Journal of Multivariate Analysis*, 11, 474–484.

DOORNIK, J. A. (1999): *Object-Oriented Matrix Programming Using Ox, 3rd ed.* Timberlake Consultants Press and Oxford, Oxford, U.K., www.nuff.ox.ac.uk/Users/Doornik.

DOUKHAN, P. (1995): *Mixing - Properties and Examples*, no. 85 in Lecture Notes in Statistics. Springer-Verlag.

DROST, F. C. (1993): "Temporal Aggregation of Time-Series," in *Econometric Analysis of Financial Markets*, ed. by J. Kaehler, and P. Kugler, pp. 11–21. Physica-Verlag, New York.

DROST, F. C., AND T. E. NIJMAN (1993): "Temporal Aggregation of GARCH Processes," *Econometrica*, 61(4), 909–927.

DURBIN, J. (1960): "The Fitting of Time Series Models," *Revue de l'Institut International de Statistique*, 28, 233.

FLORES DE FRUTOS, R., AND G. R. SERRANO (2002): "A Generalized Least Squares Estimation Method for VARMA Models," *Statistics*, 36(4), 303–316.

FRANCQ, C., R. ROY, AND J.-M. ZAKOÏAN (2003): "Goodness-of-fit tests for ARMA models with uncorrelated errors," Discussion Paper CRM-2925, Centre de recherches mathématiques, Université de Montréal.

FRANCQ, C., AND J.-M. ZAKOÏAN (1998): "Estimating Linear Representations of Nonlinear Processes," *Journal of Statistical Planning and Inference*, 68, 145–165.

GALBRAITH, J. W., A. ULLAH, AND V. ZINDE-WALSH (2000): "Estimation of the Vector Moving Average Model by Vector Autoregression," *Econometric Reviews*, 21(2), 205–219.

GALBRAITH, J. W., AND V. ZINDE-WALSH (1994): "A Simple Noniterative Estimator for Moving Average Models," *Biometrika*, 81(1), 143–155.

——— (1997): "On Simple, Autoregression-Based Estimation and Identification Techniques for ARMA Models," *Biometrika*, 84(3), 685–696.

HAMILTON, J. D. (1994): *Time Series Analysis*. Princeton University Press, Princeton, New Jersey.

HANNAN, E. J. (1969): "The Identification of Vector Mixed Autoregressive- Moving Average Systems," *Biometrika*, 57, 223–225.

——— (1971): "The Identification Problem for Multiple Equation Sysytem with Moving Average Errors," *Econometrica*, 39, 751–766.

——— (1976): "The Identification and Parameterization of ARMAX and State Space Forms," *Econometrica*, 44(4), 713–723.

HANNAN, E. J., AND M. DEISTLER (1988): *The Statistical Theory of Linear Systems*. John Wiley & Sons, New York.

HANNAN, E. J., AND L. KAVALIERIS (1984a): "A Method for Auregressive-Moving Average Estimation," *Biometrika*, 72(2), 273–280.

——— (1984b): "Multivariate Linear Time Series Models," *Advances in Applied Probability*, 16, 492–561.

——— (1986): "Regression, Autoregression Models," *Journal of Time Series Analysis*, 7(1), 27–49.

HANNAN, E. J., L. KAVALIERIS, AND M. MACKISACK (1986): "Recursive Estimation of Linear Systems," *Biometrika*, 73(1), 119–133.

HANNAN, E. J., AND J. RISSANEN (1982): "Recursive Estimation of Mixed Autoregressive-Moving-Average Order," *Biometrika*, 69, 81–94, Errata 70 (1982), 303.

——— (1983): "Errata: "Recursive Estimation of Mixed Autoregressive-Moving Average Order"," *Biometrika*, 70(1), 303.

HORN, R. G., AND C. A. JOHNSON (1985): *Matrix Analysis*. Cambridge University Press, Cambridge, U.K.

HUANG, D., AND L. GUO (1990): "Estimation of Nonstationary ARMAX Models Based on the Hannan-Rissanen Method," *The Annals of Statistics*, 18(4), 1729–1756.

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

KEATING, J. W. (2002): "Structural Inference with Long-Run Recursive Empirical Models," *Macroeconomic Dynamics*, 6(2), 266–283.

KOREISHA, S. G., AND T. M. PUKKILA (1989): "Fast Linear Estimation Methods for Vector Autoregressive Moving-Average Models," *Journal of Time Series Analysis*, 10(4), 325–339.

——— (1990a): "A Generalized Least-Squares Approach for Estimation of Autoregressive-Moving-Average Models," *Journal of Time Series Analysis*, 11(2), 139–151.

——— (1990b): "Linear Methods for Estimating ARMA and Regression Models with Serial Correlation," *Communications in Statistics, Part B -Simulation and Computation*, 19, 71–102.

——— (1995): "A Comparison Between Different Order-Determination Criteria for Identification of ARIMA Models," *Journal of Business and Economic Statistics*, 13, 127–131.

LEWIS, R., AND G. C. REINSEL (1985): "Prediction of Multivariate Time Series by Autoregressive Model Fitting," *Journal of Multivariate Analysis*, 16, 393–411.

LJUNG, G. M., AND G. E. P. BOX (1978): "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, 65(2), 297–303.

LUCEÑO, A. (1994): "A Fast Algorithm For The Exact Likelihood Of Stationary And Partially Nonstationnary Vector Autoregressive-Moving Average Processes," *Biometrika*, 81(3), 555–565.

LÜTKEPOHL, H. (1991): *Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin.

——— (1993): *Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, second edn.

LÜTKEPOHL, H., AND H. CLAESSEN (1997): "Analysis of Cointegrated VARMA Processes," *Journal of Econometrics*, 80, 223–239.

LÜTKEPOHL, H., AND D. S. POSKITT (1996a): "Consistent Estimation of the Number of Cointegration Relations in a Vector Autoregressive Model," Discussion Paper 74, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.

LÜTKEPOHL, H., AND D. S. POSKITT (1996b): "Specification of Echelon-Form VARMA Models," *Journal of Business and Economic Statistics*, 14, 69–80.

MAURICIO, J. A. (2002): "An Algorithm for the Exact Likelihood of a Stationary Vector Autoregressive-Moving Average Model," *Journal of Time Series Analysis*, 23(4), 473–486.

MCMILLIN, W. D. (2001): "The Effect of Monetary Policy Shocks: Comparing Contemporaneous versus Long-Run Identifying Restrictions," *Southern Economic Journal*, 67(3), 618–636.

NEWEY, W. K., AND K. D. WEST (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

NIJMAN, T. H., AND F. PALM (1990): "Parameter Identification in ARMA Processes in the Presence of Regular But Incomplete Sampling," *Journal of Time Series Analysis*, 11(3), 239–248.

NSIRI, S., AND R. ROY (1992): "On the Identification of ARMA Echelon-Form Models," *Canadian Journal of Statistics*, 20(4), 369–386.

NSIRI, S., AND R. ROY (1996): "Identification of Refined ARMA Echelon Form Models for Multivariate Time Series," *Journal of Multivariate Analysis*, 56, 207–231.

PALM, F., AND T. H. NIJMAN (1984): "Missing Observations in the Dynamic Regression Model," *Econometrica*, 52(6), 1415–1436.

POSKITT, D. S. (1987): "A Modified Hannan-Rissanen Strategy for Mixed Autoregressive-Moving Average Oder Determination," *Biometrika*, 74(4), 781–790.

——— (1992): "Identification of Echelon Canonical Forms for Vector Linear Processes Using Least Squares," *The Annals of Statistics*, 20, 195–215.

POSKITT, D. S., AND H. LÜTKEPOHL (1995): "Consistent Specification of Cointegrated Autoregressive Moving-Average Systems," Discussion Paper 54, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.

PUKKILA, T., S. KOREISHA, AND A. KALLINEN (1990): "The Identification of ARMA Models," *Biometrika*, 77(3), 537–548.

REINSEL, G. C., S. BASU, AND S. F. YAP (1992): "Maximum likelihood estimators in the multivariate autoregressive moving-average model from a generalized least squares viewpoint," *Journal of Time Series Analysis*, 13(2), 133–145.

SAIKKONEN, P. (1986): "Asymptotic Properties of some Preliminary Estimators for Autoregressive Moving Average Time Series Models," *Journal of Time Series Analysis*, 7, 133–155.

SHEA, B. L. (1989): "The Exact Likelihood of a Vector Autoregressive Moving Average Model," *Applied Statistics*, 38(1), 161–184.

SPLIID, H. (1983): "A Fast Estimation Method for the Vector Autoregressive Moving Average Model with Exogenous Variables," *Journal of the American Statistical Association*, 78(384), 843–849.

TIAO, G. C., AND R. S. TSAY (1983): "Consistency Properties of Least Squares Estimates of Autoregressive Parameters in ARMA Models," *The Annals of Statistics*, 11, 856–871.

——— (1989): "Model Specification in Multivariate Time Series," *Journal of the Royal Statistical Society, Series B*, 51(2), 157–213.

TSAY, R. S. (1989a): "Identifying Multivariate Time Series Models," *Journal of Time Series Analysis*, 10, 357–372.

TSAY, R. S. (1989b): "Parsimonious Parameterization of Vector Autoregressive Moving Average Models," *Journal of Business and Economic Statistics*, 7(3), 327–341.

TSAY, R. S. (1991): "Two Canonical Forms for Vector VARMA Processes," *Statistica Sinica*, 1, 247–269.

WALLIS, K. F. (1977): "Multiple time series analysis and the final form of econometric models," *Econometrica*, 45(6), 1481–1497.

ZELLNER, A., AND F. PALM (1974): "Time series analysis and simultaneous equation econometric model," *Journal of Econometrics*, 2(1), 17–54.

ZHAO-GUO, C. (1985): "The Asymptotic Efficiency of a Linear Procedure of Estimation for ARMA Models," *Journal of Time Series Analysis*, 6(1), 53–62.