

**Statistica Sinica Preprint No: SS-2016-0415.R1**

<b>Title</b>	Constrained Estimation of Causal Invertible VARMA
<b>Manuscript ID</b>	SS-2016-0415.R1
<b>URL</b>	<a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a>
<b>DOI</b>	10.5705/ss.202016.0415
<b>Complete List of Authors</b>	Anindya Roy Tucker S. McElroy and Peter Linton
<b>Corresponding Author</b>	Anindya Roy
<b>E-mail</b>	anindya@umbc.edu
Notice: Accepted version subject to English editing.	

# Constrained Estimation of Causal Invertible VARMA

Anindya Roy<sup>1,2\*</sup>, Tucker S. McElroy<sup>2</sup> and Peter Linton<sup>1</sup>

<sup>1</sup>Department of Mathematics and Statistics

University of Maryland Baltimore County, Baltimore MD 21250

<sup>2</sup>U.S. Census Bureau, Suitland, MD 20746

## Summary

We present a reparameterization of vector autoregressive moving average (VARMA) models that allows parameter estimation under the constraints of causality and invertibility. This reparameterization is accomplished via a bijection from the complicated causal-invertible parameter space to Euclidean space. The bijection facilitates computation of maximum likelihood estimators (MLE) via unconstrained optimization, as well as computation of Bayesian estimates via prior specification on the constrained space.

The proposed parameterization is connected to the Schur-stability of polynomials and the associated Stein transformation, which are often used in dynamical systems; we establish a fundamental characterization of Schur stable polynomials via a novel characterization of positive definite block Toeplitz matrices. Our results also generalize some classical results in dynamical systems.

*keywords:* Constrained estimation; Block Toeplitz matrix; Schur stability; Reparameterization

---

\*Corresponding author: Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100, anindya@umbc.edu

## 1. INTRODUCTION

The article develops a method for estimating the parameters of a general VARMA model under the constraint that the estimated process is causal and invertible. To our knowledge there are no existing procedures or software that maintain the restrictions of causality and invertibility in the estimation process (except in special cases such as the Yule-Walker for vector autoregressions). This paper fills that gap by producing a parameterization of the process that automatically maintains the constraints during estimation. Because VARMA models (particularly vector autoregression models) are ubiquitous in time series applications and because causality and invertibility are often imposed on the models (but not always enforced during the estimation process due to the complexity of the constraints), it is important to estimate VARMA processes under such constraints. That fact that VARMA models often lead to forecasting gains (Athanasopoulos and Vahid 2008, Simionescu 2013) compared to vector autoregressive models makes it important that there are available tools for fitting causal invertible VARMA models. The methods described in this work are equally applicable to fitting VAR models and VARMA models.

Recently, interest has grown in non-causal VARMA processes (Nyberg and Saikkonen 2014, Gouriroux and Jasiak, 2016) and non-Gaussian models (Breidt et. al., (1991, 2001), Lanne and Saikkonen 2013; Giurcanu, 2015; Nyberg et al., 2012). Results here describe the complement to the non-causal parameter space, implying possible applications to these topics and evaluation of the likelihood through autocovariances; non-Gaussian processes can be considered. But estimating causal invertible VARMA models with Gaussian innovations remains a topic of interest, particularly for producing long-term forecasts and understanding stability in linear dynamical systems.

A mean zero VARMA( $p, q$ ) process of dimension  $m$ , denoted by  $\{X_t\}$ , is defined by the relation

$$\Phi(B)X_t = \Theta(B)Z_t, \tag{1}$$

indexed by time  $t$ . The innovations,  $\{Z_t\}$ , satisfy  $E(Z_t) = 0$ ,  $\text{Var}(Z_t) = \Sigma$ , and  $\text{Cov}(Z_s, Z_t) = 0$  for  $s \neq t$ .  $B$  represents the back-shift operator; for any complex number  $z \in \mathbb{C}$ , the autoregressive

polynomial  $\Phi(z)$  and the moving average polynomial  $\Theta(z)$  are defined as

$$\Phi(z) = I_m - \Phi_1 z - \cdots - \Phi_p z^p, \quad (2)$$

$$\Theta(z) = I_m + \Theta_1 z + \cdots + \Theta_q z^q, \quad (3)$$

with  $m \times m$  coefficient matrices  $\Phi_1, \dots, \Phi_p$  and  $\Theta_1, \dots, \Theta_q$ , respectively. Throughout the paper  $I_m$  will denote the identity matrix of dimension  $m$ . A VARMA( $p, 0$ ) process will be referred to as a VAR( $p$ ) process (vector autoregression of order  $p$ ) and a VARMA( $0, q$ ) process will be referred to as a VMA( $q$ ) process (vector moving average of order  $q$ ). In the current investigation assume that the process is identified and the autoregressive and the moving average polynomials do not share any common factors (however, our parameterization does allow for the possibility of common factors). We understand the gains from methodology to remove common roots, such as reduced echelon form (Tsay 2014), but have not found shared roots to be a problem in our estimation. A process is weakly stationary if there exists a matrix-valued function  $\Gamma$  such that  $\Gamma(k) = E(X_t X'_{t-k})$  for  $k \in \mathbb{Z}$ .

The subclass of stationary VARMA processes that we study is the class of *causal-invertible* VARMA processes. The process is causal if (1) has exactly one stationary solution of the form  $X_t = \Psi(B)Z_t$  where  $\Psi(z) = \sum_{j=0}^{\infty} \Psi_j z^j$ ,  $z \in \mathbb{C}$  for a sequence of coefficient matrices  $\{\Psi_j, j \geq 0\}$ . Causality ensures the process is independent of future innovations, thereby allowing one to forecast ahead based on current and past observations. The VARMA( $p, q$ ) is called *invertible* if, based on (1), the innovation process  $Z_t$  can be given the representation  $Z_t = \Pi(B)X_t$  in terms of a stationary solution  $X_t$ . With  $\Pi(z) = \sum_{j=0}^{\infty} \Pi_j z^j$ ,  $z \in \mathbb{C}$ . Although spectral factorization methods (Zadrozny, 1998) can be used to render a VARMA model as causal and invertible, this has the disadvantage of numerical instability when roots are close to unit; moreover, one does not obtain an explicit parameterization of the parameter space, which interferes with the elicitation of Bayesian priors.

Likelihood based estimators often have better finite sample efficiency than moment based estimators (Fuller 1996, chap. 8), provided the assumed likelihood is approximately correctly specified. Thus, the MLE obtained by maximizing the full Gaussian likelihood of a causal invertible VARMA process (including the contribution of the initial observations) is preferable. However, due to com-

plexity of the causality and invertibility constraints, along with the highly complicated form of VARMA likelihood, MLE estimation becomes a nearly intractable problem.

There is a long literature of likelihood computation and optimization for the VARMA model, originating with univariate ARMA models. Before modern computing power, researchers often developed various approximations for the likelihood under which approximate MLE were obtained and their properties were studied (Whittle 1951, Tunnicliffe and Wilson 1973, Godolphin 1984). Some authors derived convenient algorithms for computing the VARMA likelihood (Nicholls and Hall 1979, Ansley 1988, Koreisha and Pukkila 1989, Reinsel et. al. 1992). Other authors (Mauricio 1995, Mauricio 1997, Mauricio 2002, Metaxoglou and Smith 2007) provide EM type algorithms along with a state-space formulation that makes likelihood computation considerably faster and improves convergence. However, none of the procedures guarantee that the estimated VARMA process is causal and invertible.

Along with maximum likelihood estimation, much attention has been devoted toward Bayesian vector autoregression (BVAR) and prior specification for such models. The popular choices for prior specification include those described in Litterman (1980), Doan et. al. (1984), Kadiyala and Karlsson (1997), and Sims and Zha (1998). These formulations usually have normal priors on the coefficients along with inverse wishart prior on the innovation covariance matrix. The supports of the normal priors are the entire Euclidean spaces, and hence there is positive probability that the posterior estimates may lie outside the constraint set determined by causality and invertibility. Depending on the sample size and the dimension of the parameter space, significant posterior probability may exist for estimating processes that are not causal, demonstrated in our simulations. This is highly undesirable in applications where long-term behavior of the underlying stationary system is being estimated. It is known that having prior mass outside the constrained parameter set could make Bayesian computation inefficient (Marin and Robert 2007) and in such situations some authors have advocated making parameter transformations to render the parameter restriction free (Albert 2009). Recent interest in Bayesian macroeconomics has spawned research articles in BVAR; (Koop and Korobilis 2009). Koop and Potter (2011) suggested Bayesian schemes for estimating time varying VARs under inequality restriction. However, their method does not guarantee that

the posterior is supported only on the causal set. Our methodology is supported on the causal set but does not elicit a convenient parameterization of the exact posterior distribution on the VARMA parameters. That is to be expected given the complicated nature of the causal invertible parameter space and samples from the induced posterior distribution generally appear normally distributed.

To avoid the numerical complexities and instabilities of constrained optimization or constrained prior specification, one can re-parameterize the problem via a transformation such that the new parameters are unconstrained. Parameter transformation has been successfully used in many complicated constrained estimation problems. Some examples include estimation of covariance matrices under positive-definiteness constraints (Lindstrom and Bates 1988, Leonard and Hsu 2002, Pinheiro and Bates 1996) and order-constrained parameters (Dunson and Neelon 2003). Previous attempts of such parameterization for univariate ARMA models include Wise (1956), Barndorff-Neilsen and Schou (1973), Marriott and Smith (1992), and Quenneville and McLeod (1992). However, the techniques of these papers are not easily adaptable to the vector case.

It is our contention that in complex constrained parameter problems, it is much easier to carry out inference under suitable parameter transformations. Transformation of parameters to quantities that are free (or nearly free) of constraints can present feasible solutions to complicated constrained inference problems. In this article, we derive a bijection of the constrained parameter space of a causal invertible VARMA to Euclidean space, providing a parameterization of causal invertible VARMA in terms of unconstrained parameters. The bijection goes beyond just likelihood based computation. It can be used to obtain minimum distance estimators that are constrained, and eases optimization in terms of the transformed parameters. The mapping is quite complicated, so we illustrate the first order VAR bijection in Appendix A of the supplementary material. Additional numerical illustrations as well as notes on computational aspects of the proposed algorithm are given in Appendix B in the supplement.

## 2. CAUSALITY, INVERTIBILITY, SCHUR-STABILITY AND ESTIMATION

A VARMA( $p, q$ ) process is causal if  $\det(\Phi(z)) \neq 0$ , for all  $z \in \mathbb{C}$  such that  $|z| \leq 1$  (Brockwell and Davis 1991, Theorem 11.3.1). The converse is also true under the assumption that the polynomials

$\det(\Phi(z))$  and  $\det(\Theta(z))$  do not share any common factor. For what follows it will be convenient to characterize the causal process in terms of the associated monic polynomial  $\tilde{\Phi}(z)$  defined as

$$\tilde{\Phi}(z) := z^p \Phi(z^{-1}) = I_m z^p - \Phi_1 z^{p-1} - \dots - \Phi_p.$$

A VARMA process defined by (1) is causal if  $\det(\tilde{\Phi}(z)) \neq 0$ , for all  $z \in \mathbb{C}$  such that  $|z| \geq 1$ , or equivalently the process in (1) is causal if all roots of  $\det(\tilde{\Phi}(z)) = 0$  lie within the open unit disc  $\mathcal{D} = \{z \in \mathbb{C} : |z| < 1\}$ . Similarly, define

$$\tilde{\Theta}(z) := z^q \Theta(z^{-1}) = I_m z^q + \Theta_1 z^{q-1} + \dots + \Theta_q.$$

Invertibility of the process is equivalent to the property that all roots of  $\tilde{\Theta}(z)$  lie within  $\mathcal{D}$ .

We refer to  $\tilde{\Phi}(z)$ ,  $\tilde{\Theta}(z)$  and  $\Sigma$  as the parameters of the process defined by (1) and when it is clear we interchangeably refer to the associated coefficient matrices  $\Phi = (\Phi_1, \dots, \Phi_p)$ ,  $\Theta = (\Theta_1, \dots, \Theta_q)$  and  $\Sigma$  as the parameters as well. Before describing the parameter space of a causal invertible VARMA process, we introduce further notations. Let  $\geq_L$  denote the Loewner partial ordering for symmetric matrices. Define

$$\mathcal{S}_{++}^m = \{\Sigma \in \mathcal{S}^m : \Sigma >_L 0\} \quad (4)$$

to be the set of all  $m \times m$  symmetric positive definite matrices that constitute the interior of the convex cone,  $\mathcal{S}_+^m$ , of  $m \times m$  positive semi-definite matrices in  $\mathcal{S}^m$ , the set of  $m \times m$  symmetric matrices. A matrix monic polynomial  $A(z) = z^k I_m - A_1 z^{k-1} - \dots - A_k$ , will be called *Schur-stable* if all roots of  $\det(A(z)) = 0$  lie within the unit disc  $\mathcal{D}$ . Such polynomials are common in the dynamical systems literature (Bhatia 1997; Kaszkurewicz and Bhaya 2000). Let

$$\begin{aligned} \mathfrak{S}^{m,k} = \{ & A(z) = z^k I_m - A_1 z^{k-1} - \dots - A_k : A_r \in \mathbb{R}^{m \times m}, \\ & r \geq 1, \text{ and } A(z) \text{ is Schur-stable} \} \end{aligned} \quad (5)$$

define the set of all  $m$ -dimensional Schur-stable matrix monic polynomials of degree  $k$ . Let any polynomial  $A(z) = z^k I_m - A_1 z^{k-1} - \dots - A_k$  be associated with the coefficient sequence  $A = [A_1, \dots, A_k]$ . We will define a sequence of matrices  $A = [A_1, \dots, A_k]$  to be Schur-stable provided the associated polynomial is Schur-stable. Then, the parameters  $(\Phi, -\Theta, \Sigma)$  of an  $m$ -dimensional

causal invertible VARMA( $p, q$ ) process belongs to the parameter space

$$\mathfrak{P} = \mathfrak{S}_p^m \times \mathfrak{S}_q^m \times \mathcal{S}_{++}^m. \quad (6)$$

Often for a VARMA( $p, q$ ) process the innovations are assumed to be Gaussian, i.e.  $Z_t \sim N(0, \Sigma)$ . Based on this assumption, a likelihood for the parameters  $(\Phi, -\Theta, \Sigma)$  can be written down and used for likelihood based inference. Under the assumption of second order stationarity, for any  $p \geq 0$ , define  $\underline{\Gamma}_p$  to be the covariance matrix of  $(X'_t, X'_{t-1}, \dots, X'_{t-p})'$ . The  $jk$ th block of  $\underline{\Gamma}_p$  is an  $m \times m$  matrix, given by  $\Gamma(k-j) = \Gamma'(j-k)$ , for  $1 \leq j, k \leq (p+1)$ . Assuming  $Z_t \stackrel{i.i.d.}{\sim} N(0, \Sigma)$ , a stationary likelihood for  $(\Phi, \Theta, \Sigma)$  based on a sample  $X = (X'_n, \dots, X'_1)'$  (written in the reverse order for notational consistency) is

$$\mathcal{L}(\Phi, \Theta, \Sigma) = (2\pi)^{-n/2} [\det(\underline{\Gamma}_{n-1})]^{-1/2} \exp(-0.5 X' \underline{\Gamma}_{n-1}^{-1} X). \quad (7)$$

where  $\underline{\Gamma}_{n-1}$  is a function of  $(\Phi, \Theta, \Sigma)$ . An available likelihood, evaluated through  $\underline{\Gamma}_{n-1}$ , immediately facilitates estimation. For causal invertible processes, maximum likelihood estimators of  $(\Phi, \Theta, \Sigma)$  can be obtained by maximizing the likelihood over the parameter space  $\mathfrak{P}$ , or Bayesian posterior estimates can be obtained based on priors specified in the range of  $(\Phi, \Theta, \Sigma)$ .

The Schur-stable space  $\mathfrak{S}_k^m$  is described by the roots of the matrix polynomials. Unfortunately, the roots are highly non-linear functions of parameters  $\Phi$  and  $\Theta$ , and often are implicitly defined. Thus, direct maximization of the likelihood (7) over the parameter space  $\mathfrak{P}$  is a computationally intractable problem. For Bayesian estimation that guarantees causal invertible estimates, one has to specify priors that are fully supported on  $\mathfrak{P}$  so that the posterior is supported within  $\mathfrak{P}$ . The posterior mode will belong to  $\mathfrak{P}$  and is a valid estimator. However,  $\mathfrak{P}$  is not convex and one has to be careful. If the true value is in the interior of  $\mathfrak{P}$ , the Bayesian posterior will be concentrated on an open convex set around the true value and averaging of posterior samples should be acceptable. In general, for non-convex sets means can be defined in an extrinsic manner, where the averaging is done on the unrestricted space under some suitable transformation and then mapped back to the non-convex set through the inverse transformation. In principle, one could specify a ‘‘flat’’ prior on the constrained space, but since the constraints are given in terms of the roots where as the likelihood is in terms of the coefficient matrices, implementation is difficult. Such a prior

lacks flexibility and may not guarantee propriety of the posterior. Other options include truncating general priors to the constrained space by rejecting samples that do not satisfy the constraints (Gelfand et.al 1992). Such methods can be highly inefficient for complex constraints, such as that of causality or invertibility in the VARMA setting, and inefficiency of the algorithm may increase greatly when parameters are near the boundary (Marin and Robert, 2007; Albert, 2009). To increase efficiency of computation one could project samples falling outside back to the boundary of the constrained space (Dunson and Neelon 2003) which would lead to prior mass on the boundary. In the VARMA example, having mass on the boundary means that the prior is entertaining non-stationary models for a stationary causal process. Therefore neither direct maximization of the likelihood nor direct prior specification are possible when the parameter space is  $\mathfrak{P}$ .

### 3. PARAMETERIZATION OF CAUSAL INVERTIBLE VARMA( $p, q$ )

We first establish a characterization of Schur-stable polynomials in  $\mathfrak{S}^{m,k}$  obtained in terms of positive definite block Toeplitz matrices. The characterization will help us define the bijection from  $\mathfrak{P}$  to a Euclidean space.

#### 3.1 Block Toeplitz parameterization

The following notation will be used hereafter. For  $j \geq 1$ , define  $\underline{U}_j$  to be a symmetric block Toeplitz matrix of order  $j$ :

$$\underline{U}_j = \begin{pmatrix} U(0) & U(1) & \cdots & U(j) \\ U(1)' & U(0) & \cdots & \cdot \\ \vdots & \ddots & \ddots & \vdots \\ U(j)' & \cdots & U(1)' & U(0) \end{pmatrix} \quad (8)$$

where  $U(0), U(1), \dots, U(j)$  are arbitrary  $m \times m$  matrices and  $U(0) \in \mathcal{S}^m$ . Note that  $\underline{U}_0 = U(0)$ . For  $j \geq 1$ . We will take advantage of the following nested representations of  $\underline{U}_j$  in terms of  $\underline{U}_{j-1}$ : the lower representation given by

$$\underline{U}_j = \begin{pmatrix} U(0) & \xi_j' \\ \xi_j & \underline{U}_{j-1} \end{pmatrix}, \quad (9)$$

and the upper representation given by

$$\underline{U}_j = \begin{pmatrix} \underline{U}_{j-1} & \kappa_j \\ \kappa'_j & U(0) \end{pmatrix}. \quad (10)$$

Here  $\xi'_j = (U(1), \dots, U(j))$  and  $\kappa'_j = (U(j)', \dots, U(1)')$ . Set  $\xi_0$  and  $\kappa_0$  equal to zero matrices. The Schur complements of  $\underline{U}_{j-1}$  in  $\underline{U}_j$  in the two representations (9) and (10) are

$$C_j = U(0) - \xi'_j \underline{U}_{j-1}^{-1} \xi_j, \quad (11)$$

$$D_j = U(0) - \kappa'_j \underline{U}_{j-1}^{-1} \kappa_j. \quad (12)$$

Define  $C_0 = D_0 = U(0)$ . Let  $\mathfrak{T}^{m,k}$  denote the set of  $m(k+1) \times m(k+1)$  symmetric block Toeplitz matrices with  $m$ -dimensional blocks, i.e.

$$\mathfrak{T}^{m,k} = \{\underline{U}_k \in \mathcal{S}^{m(k+1)} : \underline{U}_k \text{ is in the form (8)}\}$$

Also define  $\mathfrak{T}_{++}^{m,k}$  to be the subset of  $\mathfrak{T}^{m,k}$  comprising the positive definite block Toeplitz matrices of order  $k$  and  $m$ -dimensional blocks.

**Theorem 1.** *An  $m$ -dimensional matrix polynomial  $A(z) = I_m z^k - A_1 z^{k-1} \dots - A_k$  is Schur-stable if and only if there exists  $\underline{U}_k \in \mathfrak{T}_{++}^{m,k}$  such that the coefficients  $A = [A_1, \dots, A_k] \in \mathbb{R}^{m \times mk}$  satisfy the Yule-Walker relation  $A = \xi'_k \underline{U}_{k-1}^{-1}$ .*

**Remark 1.** This result is suggested in the scalar case, where AR parameters can be computed in terms of partial autocorrelations; as in Quenneville and McLeod (1992). One can parameterize each partial autocorrelation to be a number in  $(-1, 1)$ , and implicitly describe the entire parameter space. The challenge in the vector case, is determining how to parameterize a correlation matrix to fully describe the parameter space.

Theorem 1 presents a way of parameterizing Schur-stable polynomials via positive definite block Toeplitz matrices. Other representations of block Toeplitz matrices are given in (Constantinescu 1986, Delsargte et.al. 1979). The next theorem describes block Toeplitz matrices,  $\mathfrak{T}_{++}^{m,k}$ , and therefore  $\mathfrak{S}^{m,k}$  in terms of simpler objects which lend themselves conveniently to optimization and prior specification.

**Theorem 2.** A block Toeplitz matrix  $\underline{U}_k \in \mathfrak{T}^{m,k}$  is positive definite if and only if the associated Schur complement sequence  $C_j = U(0) - \xi_j' \underline{U}_{j-1}^{-1} \xi_j$  satisfies  $C_0 \geq_L C_1 \geq_L \cdots \geq_L C_k >_L 0$ .

**Remark 2.** The characterization allows the successive difference of the Schur complements,  $C_{i-1} - C_i$  to be nonnegative definite. For implementation we will use the differences as parameters to be further simplified in terms of real parameters. To this end, we will restrict to cases where the differences are strictly positive definite, and let the non-negative definite cases be approximated by the positive definite parameters as limiting values. From a practical vantage point there is an advantage in dealing with the positive definite parameterization.

Let  $C_k = M$  be fixed and fully specified. Let  $V_j = C_{j-1} - C_j$  for  $1 \leq j \leq k$ . Following the proof of Theorem 2,

$$V_j = \left( U(j) - \xi_{j-1}' \underline{U}_{j-2}^{-1} \kappa_{j-1} \right) D_{j-1}^{-1} \left( U(j) - \xi_{j-1}' \underline{U}_{j-2}^{-1} \kappa_{j-1} \right)', \quad (13)$$

which has the solution

$$U(j) = \xi_{j-1}' \underline{U}_{j-2}^{-1} \kappa_{j-1} + V_j^{1/2} Q_j D_{j-1}^{1/2} \quad (14)$$

for some orthogonal matrix  $Q_j$ . Once  $\underline{U}_j$  and  $V_j$  are specified, the orthogonal matrix is given by

$$Q_j = V_j^{-1/2} [U(j) - \xi_{j-1}' \underline{U}_{j-2}^{-1} \kappa_{j-1}] D_{j-1}^{-1/2} \quad (15)$$

Here  $V_j^{1/2}$  and  $D_{j-1}^{1/2}$  are symmetric square roots of  $V_j$  and  $D_{j-1}$ , respectively. Then (14) defines the key recursion equation that allows one to solve for  $U(j)$ ,  $j = 0, \dots, k$  iteratively, once the positive definite matrices  $M, V_1, \dots, V_k$  and the orthogonal matrices  $Q_1, \dots, Q_k$  have been specified. At the  $j$ th stage, all quantities on the right side of (14) are known; thus  $U(j)$  and  $\underline{U}_j$  can be computed. Subsequently  $\xi_j, \kappa_j$  and  $D_j$  are obtained from  $\underline{U}_j$  through (11) and (12), and then used in the next  $(j + 1)$ th iteration. The telescoping sum  $C_0 = \sum_{j=1}^k V_j + M$  allows initialization of the algorithm with  $U(0) = C_0$ .

**Algorithm [VQ]:** Algorithm for computing  $A(z)$  from  $V_1, \dots, V_k, Q_1, \dots, Q_k$

The following steps, called Algorithm [VQ], describe the algorithm for computing the block Toeplitz matrix  $\underline{U}_k$  and hence  $A$ , the coefficients of the Schur-stable polynomial  $A(z)$  from the pre-parameters  $V_1, \dots, V_k, Q_1, \dots, Q_k$ .

1. Set  $U(0) = C_0 = M + \sum_{j=1}^k V_j$ .
2. Compute  $U(1)' = V_1^{1/2} Q_1 U(0)^{1/2}$  and obtain  $\underline{U}_1$ .
3. Compute  $\kappa_1, \xi_1, D_1$  based on  $\underline{U}_1$ . Here  $\kappa_1 = \xi_1' = U(1)$  and  $D_1 = U(0) - U(1)'U(0)^{-1}U(1)$ .
4. Compute  $U(2) = \xi_1' U(0)^{-1} \kappa_1 + V_2^{1/2} Q_2 D_1^{1/2}$  and obtain  $\underline{U}_2$  from  $U(0), U(1), U(2)$ .
5. Obtain  $\kappa_2, \xi_2, D_2$  and iterate using (14).
6. Once  $U(0), U(1), \dots, U(k)$  and hence  $\underline{U}_k$  have been obtained, compute  $A$  using the Yule-Walker relation  $A = \xi_p' \underline{U}_{k-1}^{-1}$ .

The inverse algorithm is obtained by noting that  $\text{Vec}(\underline{U}_{k-1}) = (I - \tilde{A} \otimes \tilde{A})^{-1} \text{Vec}(\tilde{M})$  and  $\xi_k' = A \underline{U}_{k-1}$  where  $I$  is the identity of dimension  $(mk \times mk)$ ,  $\tilde{A}$  is defined as

$$\tilde{A} = \begin{pmatrix} A_1 & A_2 & \cdots & A_{k-1} & A_k \\ I_m & 0 & \cdots & 0 & 0 \\ \vdots & \cdots & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & I_m & 0 \end{pmatrix} \quad (16)$$

and  $\tilde{M}$  is  $\tilde{A}$  for  $(A_1, \dots, A_k) = (M, 0, \dots, 0)$ .

### 3.2 Generalized Stein transformation

For any square matrix  $A \in \mathbb{R}^{m \times m}$  and any symmetric matrix  $U \in \mathcal{S}^m$  define the transformation  $S(A, U) : \mathbb{R}^{m \times m} \times \mathcal{S}^m \rightarrow \mathcal{S}^m$ , by

$$S(A, U) = U - AUA'. \quad (17)$$

For any fixed  $A \in \mathbb{R}^{m \times m}$ , the  $A$ -Section of the transformation, defined by  $S_A(U) = S(A, U)$ , is a map from  $\mathcal{S}^m \rightarrow \mathcal{S}^m$ . It is known as the *Stein* transformation with respect to  $A$  and has been extensively studied in the dynamical systems literature in relation to stability of discrete dynamical systems. Stein (1952) showed that there exists a  $U \in \mathcal{S}_{++}^m$  such that  $S(A, U) \in \mathcal{S}_{++}^m$  if and only if  $A \in \mathfrak{S}_1^m$ . Stein's result implies that one could characterize  $\mathfrak{S}_1^m$  in terms of matrices in  $\mathcal{S}_{++}^m$ . For

any  $M \in \mathcal{S}_{++}^m$ , the pre-image  $A_M(U) = \{A : S(A, U) = M\}$  is non-empty if and only if  $U \geq_L M$ , and need not be a singleton set.

We define a generalization of the Stein transformation on the set of positive definite block Toeplitz matrices, allowing characterization of stability properties in high order monic matrix polynomials. Fix a set of coefficient matrices  $A = [A_1, \dots, A_k] \in \mathbb{R}^{m \times mk}$  associated with a polynomial  $A(z) = I_m z^k - A_1 z^{k-1} - \dots - A_k$ , and a symmetric matrix  $\underline{U} \in \mathcal{S}^{mk}$  with  $U_{11}$  as the upper left  $m \times m$  block of  $\underline{U}$ . Define the *Generalized Stein Transformation*  $\tilde{S}_k(A, \underline{U}) : \mathbb{R}^{m \times mk} \times \mathcal{S}^{mk} \rightarrow \mathcal{S}^{mk}$  by

$$\tilde{S}_k(A, \underline{U}) = \underline{U} - \tilde{A} \underline{U} \tilde{A}', \quad (18)$$

in other words,  $\tilde{S}_k(A, \underline{U}) = S(\tilde{A}, \underline{U})$ . Define  $S_k(A, \underline{U}) = U_{11} - \underline{A} \underline{U} \underline{A}'$  to be the upper left  $m \times m$  block of  $\tilde{S}_k(A, \underline{U})$ . The Generalized Stein Transformation reduces to the Stein transformation for the case  $k = 1$ . Analogous to Stein (1952), one can characterize Schur-stability of  $A(z)$  from properties of the transformation.

**Theorem 3.** *A matrix polynomial  $A(z) = z^k - A_1 z^{k-1} - \dots - A_k$  with coefficients  $A = [A_1, \dots, A_k]$  is Schur-stable if and only if there exists a positive definite block Toeplitz matrix  $\underline{U}_{k-1} \in \mathfrak{T}_{++}^{m, k-1}$ , such that the Generalized Stein Transformation  $\tilde{S}_k(A, \underline{U}_{k-1}) \in \mathcal{S}_+^{mk}$  and  $S_k(A, \underline{U}_{k-1}) \in \mathcal{S}_{++}^m$ .*

**Remark 3.** In general, for self-dual cones in finite dimensional Hilbert space with a Euclidean Jordan algebra, characterization of Stein-type operators can be done (Schneider 1965). However, due to the special structure of  $\tilde{A}$  a more refined result on positivity of the generalized transformation can be obtained.

**Remark 4.** The quantity  $S_k(A, \underline{U}_{k-1})$  is precisely the Schur-complement  $C_k$  under the conditions of Theorem 1. Also, from the proof of Theorem 3 it is clear that if  $\underline{U}_{k-1}$  satisfying the conditions of Theorem 3 exists, then necessarily  $\tilde{S}_k(A, \underline{U})$  will be of the form

$$\tilde{S}_k(A, \underline{U}) = \begin{pmatrix} S_k(A, \underline{U}) & 0 \\ 0 & 0 \end{pmatrix}.$$

### 3.3 The role of $M$

For any fixed known  $M \in \mathcal{S}_{++}^m$ , consider the pre-image

$$A_M(\underline{U}_{k-1}) = \{A : S_k(A, \underline{U}_{k-1}) = M\}.$$

The set is non-empty: Theorem 2 describes the construction of a positive definite block Toeplitz matrix  $\underline{U}_k$  with  $C_k = M$  for general  $M$ . Then, by Theorem 1, any  $A$  of the form  $A = \xi_k' \underline{U}_{k-1}^{-1}$  will be a member of the pre-image. Additionally,

$$\mathfrak{S}^{m,k} = \bigcup_{\underline{U}_{k-1} \in \mathfrak{S}_{++}^{m,k}} A_M(\underline{U}_{k-1}).$$

The result is established by noting that given  $A \in \mathfrak{S}^{m,k}$  and  $M \in \mathcal{S}_{++}^m$ , one can construct a causal VAR( $k$ ) model with  $A$  as the coefficients and  $M$  as the innovation variance and then  $\underline{U}_{k-1} = \underline{\Gamma}_{k-1}$  will satisfy the Generalized Stein Transformation. This reinforces the point that the class of Schur-stable polynomials can be parameterized by the class of positive definite block Toeplitz matrices, which in turn is accomplished using Theorem 2.

### 3.4 Further reparametrization

Since the objective is to map the constrained space  $\mathfrak{P}$  to a Euclidean space, further parametrization of  $(V, Q, \Sigma)$  in terms of unrestricted real numbers is desirable. For the positive definite matrices, one could use the forms described in Lindstrom and Bates (1988) or Leonard and Hsu (2002). Alternatively, other forms of decomposition given in terms of eigenvalues and eigenvectors can be pursued. The parameterization of positive definite matrices in terms of their Cholesky decomposition given in Lindstrom and Bates (1988) is particularly useful. The specific form for the  $m \times m$  positive definite matrix  $V$  is

$$V = LDL' \tag{19}$$

where  $L$  and  $D$  are given by

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ l_{m,1} & l_{m,2} & \cdots & 1 \end{pmatrix}, \quad D = \begin{pmatrix} e^{d_1} & 0 & \cdots & 0 \\ 0 & e^{d_2} & \cdots & 0 \\ \vdots & \cdots & \ddots & 0 \\ 0 & 0 & \cdots & e^{d_m} \end{pmatrix}.$$

The representation, which has  $\binom{m+1}{2}$  free parameters of the positive definite matrix  $V$  in terms of the  $\binom{m}{2}$  unconstrained real numbers in  $l = (l_{2,1}, \dots, l_{m,m-1})$  and  $m$  unconstrained real numbers in  $d = (d_1, \dots, d_m)$ , is a bijection.

Let  $O(m)$  denote the group of  $m \times m$  orthogonal matrices. Let the special orthogonal group  $SO(m)$ , be the set of matrices in  $O(m)$  with determinant equal to one. The orthogonal matrix  $Q \in O(m)$  is neither suitable for optimization nor for prior specification. Any element in  $O(m)$  can be connected to one in  $SO(m)$  through a single Householder reflection. Let  $E_\delta = I_m - 2\delta e_1 e_1'$ , where  $\delta \in \{0, 1\}$ , and  $e_1 = (1, 0, \dots, 0)'$ . Then any element  $Q \in O(m)$  can be viewed as  $Q = E_\delta R$ , for some  $\delta \in \{0, 1\}$  and some  $R \in SO(m)$ . Other options describing  $SO(m)$  include Givens rotations (angles) or the Cayley representation, which excludes any  $R \in SO(m)$  with an even number of negative one eigenvalues. Gaillier (2013) provides a complete parameterization of  $SO(m)$  given by  $R = [(I_m - S)(I_m + S)^{-1}]^2$  for some skew-symmetric matrix  $S$ . Every orthogonal matrix  $Q \in O(m)$  can be given a *modified Cayley form*

$$Q = E_\delta [(I_m - S)(I_m + S)^{-1}]^2, \quad (20)$$

for some skew-symmetric matrix  $S$  and some reflection  $E_\delta$ . The inverse transformation that determines  $S$  from  $R$  can be defined in an injective manner following Proposition 1.3 in Gaillier (2013). Let  $S = 2(I_m + R^{1/2})^{-1} - I_m$  be the inverse image of  $R$  for the transformation (20), where  $R^{1/2}$  is the unique square root of  $R$  in  $SO(m)$  without any negative one eigenvalues. The matrix  $R^{1/2}$  is defined following the normal form algorithm in Proposition 1.3 in Gaillier (2013). The mapping of  $Q$  to the  $\frac{m(m-1)}{2}$  unrestricted elements in  $s = (s_{21}, \dots, s_{m,m-1})$  and the binary parameter  $\delta$  is a bijection. Thus, if we define the map

$$\tau(V, Q) = (l, d, s, \delta) \quad (21)$$

then  $\tau$  is a bijection that maps  $\mathcal{S}_{++}^m \times O(m)$  to  $\mathbb{R}^{m(m+1)/2} \times \{0, 1\}$ .

### 3.5 Family of bijections

The Cholesky transformation (19) and the transformation (20) provide the final steps in mapping the coefficient matrices of a Schur-stable polynomial to unrestricted real numbers. The mapping depends on the positive definite matrix  $M$  chosen to define the first step of the mapping from the Schur-stable space to the space of positive definite block Toeplitz matrices. The family of mappings from the Schur-stable space to the real numbers is indexed by positive definite matrices. The following theorem shows that for each  $M > 0$  the map is bijective, and hence there is no loss of information in the proposed parameterization.

**Theorem 4.** *For each  $m \times m$  positive definite matrix  $M$ , there exists a bijection  $f_M : \mathfrak{S}_k^m \mapsto \mathbb{R}^{km^2} \times \{0, 1\}^k$ . The bijection can be decomposed as  $f_M = \tau \circ \psi_M \circ \phi_M$  where*

$$\tau : (\mathcal{S}_{++}^m)^k \times O(m)^k \mapsto \mathbb{R}^{km^2} \times \{0, 1\}^k,$$

$$\psi_M : \mathfrak{T}_{++}^{m,k}(M) \mapsto (\mathcal{S}_{++}^m)^k \times O(m)^k$$

$$\phi_M : \mathfrak{S}_k^m \mapsto \mathfrak{T}_{++}^{m,k}(M).$$

All are bijections and can be described as follows:

1. For each  $A \in \mathfrak{S}_k^m$ ,  $\phi_M(A)$  is  $\underline{U}_k$ , the variance matrix of  $Y' = (Y'_1, \dots, Y'_k)'$  where  $\{Y_t\}$  is a VAR( $k$ ) process with coefficient matrices given by  $A$  and innovation variance equal to  $M$ .
2. For each  $\underline{U}_k \in \mathfrak{T}_{++}^{m,k}(M)$ ,  $\psi_M(\underline{U}_k) = (V_1, \dots, V_k, Q_1, \dots, Q_k)$  where  $V_j, Q_j, j = 1, \dots, k$ , are given iteratively by equations (13) and (15).
3. For each set of  $(V_1, \dots, V_k, Q_1, \dots, Q_k) \in (\mathcal{S}_{++}^m)^k \times O(m)^k$  the map  $\tau$  is

$$\tau(V_1, \dots, V_k, Q_1, \dots, Q_k) = (\underline{l}, \underline{d}, \underline{s}, \underline{\delta})$$

The quantities  $\underline{l}$  and  $\underline{d}$  are the parameters of the Cholesky transformation (19) for the set of all  $V_j$ .  $\underline{s}$  and  $\underline{\delta}$  are the parameters of the transformation (20) for the set of all  $Q_j$ .

#### 4. NUMERICAL COMPUTATION

Even without the constraints of causality and invertibility, numerical computation and estimation for autoregressive moving average models with nonzero moving average components have traditionally been quite challenging. This difficulty in computation is due to the non-linearity of the likelihood with respect to the moving average parameters. The parameterization described in the previous section provides a feasible way for carrying out likelihood based estimation for parameters of a causal invertible VARMA( $p, q$ ) process. Before proceeding we describe our convention for denoting the parameters associated with the autoregressive part and the moving average parts, respectively.

The matrix valued parameters associated with the autoregressive part will be denoted as  $(V_1^1, \dots, V_p^1)$  and  $(Q_1^1, \dots, Q_p^1)$  whereas those associated with the moving average part will be denoted as  $(V_1^2, \dots, V_q^2)$  and  $(Q_1^2, \dots, Q_q^2)$ , respectively. The innovation variance will be denoted by  $\Sigma$ . The matrix  $M$ , needed to start the parameterization of the Schur-stable polynomials will be taken as the identity matrix  $I_m$  for both the autoregressive and the moving average polynomials. Following notations developed earlier, the scalar valued parameters associated with Cholesky decomposition of the  $V$  matrices for the autoregressive part will be denoted as  $\underline{l}^{1,j} = (l_{2,1}^{1,j}, \dots, l_{m,m-1}^{1,j})$  and  $\underline{d}^{1,j} = (d_1^{1,j}, \dots, d_m^{1,j})$  for  $j = 1, \dots, p$ , and as  $\underline{l}^{2,j} = (l_{2,1}^{2,j}, \dots, l_{m,m-1}^{2,j})$  and  $\underline{d}^{2,j} = (d_1^{2,j}, \dots, d_m^{2,j})$ ,  $j = 1, \dots, q$  for the moving average part. Similarly the scalar parameters associated with the orthogonal matrices will be denoted as  $\underline{s}^{1,j} = (s_{2,1}^{1,j}, \dots, s_{m,m-1}^{1,j})$  and  $\underline{\delta}^{1,j}$ ,  $j = 1, \dots, p$ , for the autoregressive part and as  $\underline{s}^{2,j} = (s_{2,1}^{2,j}, \dots, s_{m,m-1}^{2,j})$  and  $\underline{\delta}^{2,j}$  for  $j = 1, \dots, q$  for the moving average part. The Cholesky parameters associated with the innovation variance are written as  $\underline{l}^0 = (l_{2,1}^0, \dots, l_{m,m-1}^0)$  and  $\underline{d}^0 = (d_1^0, \dots, d_m^0)$ . All together the real scalar parameters are then  $(\underline{l}, \underline{d}, \underline{s})$  where  $\underline{l} = (\underline{l}^{1,1}, \dots, \underline{l}^{1,p}, \underline{l}^{2,1}, \dots, \underline{l}^{2,q}, \underline{l}^0)$  and  $\underline{d}, \underline{s}$  are similarly defined.

##### 4.1 Maximum likelihood

The standard optimization methods are either gradient based stepping algorithms or direct search algorithms. Gradient based methods for the VARMA model may have difficulty due to the non-

linear nature of the likelihood. Direct search methods may fail due to the high number of parameters. For maximum likelihood computation it is convenient to use the Cholesky form for the positive definite part of the parameterization. The parameters  $\underline{l}$ ,  $\underline{d}$  and  $\underline{s}$ , are unrestricted real numbers. For a VARMA( $p, q$ ) the  $\delta$  parameters  $\underline{\delta} = (\underline{\delta}^{1,1}, \dots, \underline{\delta}^{1,p}, \underline{\delta}^{2,1}, \dots, \underline{\delta}^{2,q})$  lie in  $\{0, 1\}^{p+q}$  providing  $2^{p+q}$  possible values for the parameters. Suppose the VARMA likelihood is written as  $\mathcal{L}(\Phi, \Theta, \Sigma)$ , which is a function of the parameters  $\underline{l}$ ,  $\underline{d}$ ,  $\underline{s}$  and  $\underline{\delta}$  via the bijection  $f_M$  given in Theorem 4. The maximum likelihood estimators of the scalar parameters are defined as

$$(\hat{l}, \hat{d}, \hat{s}, \hat{\delta}) = \arg \max_{\delta \in \{0,1\}^{p+q}} \arg \max_{\underline{l}, \underline{d}, \underline{s}} \mathcal{L}(\Phi, \Theta, \Sigma),$$

and hence that of the matrix parameters as

$$(\hat{\Phi}, \hat{\Theta}, \hat{\Sigma}) = f_M^{-1}(\hat{l}, \hat{d}, \hat{s}, \hat{\delta}).$$

For all computations we use  $M = I_m$ . When  $2^{p+q}$  is moderate, a profile approach is recommended, computing maximizers for each value of  $\delta$  and then taking the largest likelihood value. In practice, initial estimates on the  $2^p \delta$  values associated with the autoregression can be fixed and only the  $2^q$  values might need be considered. We initialize with a crude but fast estimate of  $\Phi$  and  $\Theta$ , and if the estimates do not satisfy the constraints they are shrunk towards the constrained space following the algorithm [SHRINK] given in the supplementary material. A fast consistent estimator is provided by the Hannan and Rissanen (1982) algorithm.

#### 4.2 Bayesian Prior Specification and Computation

For the VAR model, Normal-Inverse Wishart (NIW) distributions are popular choices for priors on the coefficient matrices and the innovation variance. The Minnesota prior (Litterman 1980) also follows normal specifications for coefficients of individual equations in the VAR model. Prudent choice of the hyperparameters in these specifications can lead to better forecasting properties for the BVAR. However, none of the current prior choices restrict the prior, and thereby the posterior, to the causal invertible space. More specifically, as described in Section 3, the posterior probability of estimating a model with unit root or roots outside the unit circle remain significant when the sample size is small.

The parameterization given in this article restricts the prior to the causal invertible space for a general VARMA model. An advantage of the transformation is that standard normal priors can be used for all the scalar parameters (including the reflection  $\delta$  by writing it as  $I(z > 0)$  for a standard normal variate  $z$ ). For prior specification one could directly use the parameterization based on matrices or use those based on scalar parameters. For the positive definite part of the matrix valued parameterization there are several options with obvious prior choices being Wishart or Inverse-Wishart. For the orthogonal matrices belonging to  $O(m)$  the choices are more limited. For prior specification on  $SO(m)$  one could specify the uniform prior which will be proper due to compactness of  $SO(m)$ . Other options for direct prior specification on  $SO(m)$  include (Chikuse 2003), the Bingham-von Mises-Fisher (BMF) distribution (Hoff 2009), and those involving the Langevin density on  $SO(m)$  (Chiuso et.al. 2008).

## 5. SIMULATION AND DATA EXAMPLES

In this section we present simulation results to illustrate the performance of MLEs and constrained Bayesian estimators computed using the proposed parameterization. We compare with a Yule-Walker estimator, which is causal. Lastly a dataset is analyzed with a causal invertible VARMA model.

### 5.1 Two-dimensional VAR(1)

We consider the simplest model in the VARMA( $p, q$ ) class, namely a first order two-dimensional vector autoregression:

$$X_t = \begin{pmatrix} \Phi_{11} & 0 \\ 1 & .8 \end{pmatrix} X_{t-1} + Z_t, \quad (22)$$

where  $Z_t \stackrel{iid}{\sim} N(0, I_2)$ . We vary the upper diagonal entry  $\Phi_{11}$ , because it is also one of the eigenvalues of  $\Phi_1$ . The values of  $\Phi_{11}$  are chosen from the set  $\{-.95, -.95, -.9, -.8, -.6, -.4, -.2, 0, .2, .4, .6, .8, .9, .95, .99\}$  to examine performance over a varying parameter space. Maximum likelihood estimation is done using the *optim* L-BFGS-B function in R. The initial values of the pre-parameters are chosen to be those calculated from the Yule-Walker estimate. We use box constraints with upper

and lower bounds for the real parameters. The bounds are  $\pm 1e+30$  for the  $l$  and  $s$  pre-parameters and  $\pm 1e+10$  for the  $d$  parameters. A tighter bound for the  $d$  parameter prevents the iterations from generating near singular values of the  $V$  matrix. For Bayesian estimation,  $N(0, 5)$  prior is specified for all the pre-parameters except  $\delta$ , which is assigned a *Bernoulli*(0.5) prior. The Bayesian updates are done with Metropolis random walk for the real parameters and via independent sampling with a jump distribution of *Bernoulli*(0.5) for  $\delta$ . Metropolis chains reported are from length 20,000, with a burn-in of 5000.

Let the Mean Square Error for the VAR( $p$ ) polynomial coefficient matrices be defined as  $N^{-1} \sum_{j=1}^N \sum_{k=1}^p \|\hat{\Phi}_j - \Phi\|^2$ , where  $\|\cdot\|$  is the Frobenius norm of a matrix and  $\hat{\Phi}_{j,k}$  is the  $j$ th Monte Carlo estimate of  $\Phi_k$ . For a single entry in a coefficient matrix the Monte Carlo MSE is defined as the average square distance of the estimated value to the true value where the average is over  $N = 1000$  Monte Carlo replications. We report the square root of the MSE (RMSE) as a function of  $\Phi_{11}$  and compared the overall RMSE with the asymptotic value obtained based on the asymptotic variance of the individual entries of  $\Phi$ . The RMSE is plotted as a function of  $\Phi_{11}$  in Figure 1. From the figure we see that the likelihood based estimators are performing better than the Yule-Walker estimator, particularly when the largest root is close to unity in magnitude. All three estimators are causal and have similar bias but the gain in efficiency for the MLE and Bayes estimates are largely due to reduction in variance. In terms of the overall RMSE, the likelihood based estimators are nearly twice as efficient as the Yule-Walker estimator when the largest root of the VAR coefficient is close to one.

## 5.2 Data Example

The data example is based on a series analyzed by Tsay (2014). The data is from the Federal Reserve Bank of St. Louis (FRED Data) and comprises two series: monthly personal consumption expenditure (PCE) and disposable personal income (DSPI) from January 1959 to January 2015. The series are in billions of dollars and are seasonally adjusted. As argued by Tsay (2014), a VARMA(3,1) model with Gaussian innovations fits the log differenced series reasonably well. We model the differenced log-scale series multiplied by 100, modeling the growth rates in percentages.

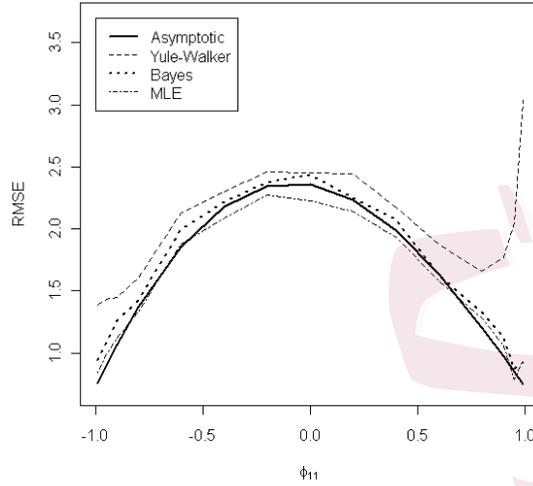


Figure 1: Overall RMSE,  $nN^{-1} \sum_{j=1}^N \|\hat{\Phi} - \Phi\|^2$ , for different estimators of  $\Phi$  compared with the corresponding asymptotic value. The RMSE is plotted as a function of  $\Phi_{11}$  in model (22).

The series in this scale are plotted in Figure 2. We compare estimated VARMA(3,1) parameters of MLE results from our transformation approach with those obtained using the MTS package in Tsay (2014) (denoted by MTS).

Table 1: Estimated parameters using the MTS method and the MLE using the proposed parameterization for the FRED Data. The standard errors for the estimated coefficients are given in parenthesis under the estimates.

Method	$\Phi_{11}^{(1)}$	$\Phi_{21}^{(1)}$	$\Phi_{12}^{(1)}$	$\Phi_{22}^{(1)}$	$\Phi_{11}^{(2)}$	$\Phi_{21}^{(2)}$	$\Phi_{12}^{(2)}$	$\Phi_{22}^{(2)}$
MTS	0.472 (0.093)	0.445 (0.067)	0.272 (0.130)	0.356 (0.115)	0.058 (0.050)	0.135 (0.040)	0.098 (0.067)	-0.062 (0.056)
MLE	0.457 (0.060)	0.411 (0.049)	0.284 (0.104)	0.357 (0.092)	0.060 (0.037)	0.139 (0.026)	0.116 (0.052)	-0.037 (0.038)
Method	$\Phi_{11}^{(3)}$	$\Phi_{21}^{(3)}$	$\Phi_{12}^{(3)}$	$\Phi_{22}^{(3)}$	$\Theta_{11}^{(1)}$	$\Theta_{21}^{(1)}$	$\Theta_{12}^{(1)}$	$\Theta_{22}^{(1)}$
MTS	-0.017 (0.045)	0.191 (0.044)	0.091 (0.065)	-0.070 (0.059)	-0.675 (0.084)	-0.345 (0.060)	-0.195 (0.121)	-0.642 (0.110)
MLE	-0.012 (0.030)	0.186 (0.030)	0.096 (0.046)	-0.069 (0.044)	-0.650 (0.054)	-0.311 (0.043)	-0.199 (0.092)	-0.642 (0.086)

The two methods yield similar estimates with comparable standard errors, with the MLE gener-

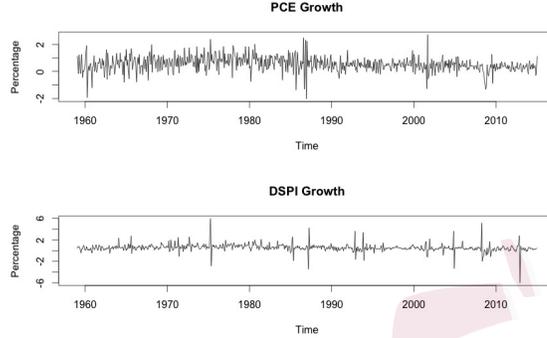


Figure 2: Plots of log differenced monthly personal consumption expenditure (PCE) and disposable personal income (DSPI)

ally having smaller standard errors. After adjusting for multiple comparison, the list of coefficients that are deemed statistically significantly different from zero are identical under the two methods. The estimates of the innovation variance were very similar under the two methods (along with nearly identical standard errors, not reported here). The estimated innovation variances were

$$\hat{\Sigma}_1^{MTS} = \begin{pmatrix} 0.269 & 0.085 \\ 0.085 & 0.458 \end{pmatrix}, \quad \hat{\Sigma}_1^{MLE} = \begin{pmatrix} 0.262 & 0.090 \\ 0.090 & 0.460 \end{pmatrix},$$

for the two methods. For the proposed method, the standard errors of the estimated parameters in the reparametrized form are obtained using numerical Gradient and Hessian approximations obtained from the numerical optimization routine (R optim). Since the method is based on a transformation, the standard error of the original parameters in terms of those for the transformed parameters are obtained by using a numerical linearization of the transformation and applying the delta method.

### 5.3 Probability of Estimating Noncausal Model Using Unrestricted Methods

To further motivate our parameterization, we demonstrate how models from unrestricted methods with a root close to the boundary for the  $\Phi$  polynomial frequently estimate non-causal models. To demonstrate this phenomenon we simulate one thousand datasets generated from multiple stationary 2 dimensional VAR(1) processes and fit using the R function `rfvar3` used by Sims (2010). For

the VAR(1) setting, 1000 series of length  $n$  are generated with an autoregressive parameter using a local-to-unity parameterization, with the largest root and first element of  $\Phi$  equal to  $(1 - 1/n)$  for data of length  $n$ . The data generating process has Gaussian error and is parameterized by:

$$\Phi = \begin{pmatrix} (1 - 1/n) & 0 \\ a_{21} & a_{22} \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The proportion of cases that fits were non causal for various  $n$ ,  $a_{21}$  and  $a_{22}$  are given in Table 2.

Table 2: Proportion of simulated non-causal estimates reported by rfvar3

$a_{21} = .1$	$n = 50$	$n = 100$	$n = 500$	$a_{21} = 1.0$	$n = 50$	$n = 100$	$n = 500$
$a_{22} = 0.8$	0.193	0.096	0.022	$a_{22} = 0.8$	0.197	0.090	0.020
$a_{22} = 0.9$	0.298	0.127	0.025	$a_{22} = 0.9$	0.262	0.128	0.040
$a_{22} = .95$	0.367	0.186	0.039	$a_{22} = .95$	0.354	0.193	0.051

One thousand datasets are also generated from a causal invertible 2 dimensional VARMA(1,1) processes and fit using the MTS package VARMA function. The VARMA(1,1) model used in the simulation is given in (23). The moving average parameter is fixed for all cases. The error was Gaussian. The proportion of cases for which the estimate was non-causal is given in Table 3.

$$\Phi = \begin{pmatrix} (1 - 1/n) & 0 \\ a_{21} & a_{22} \end{pmatrix} \quad \Theta = \begin{pmatrix} .5 & .2 \\ .2 & .5 \end{pmatrix} \quad \Sigma = \begin{pmatrix} .5 & 0 \\ 0 & .5 \end{pmatrix} \quad (23)$$

Table 3: Proportion of simulated non-causal estimates reported by MTS

$a_{21} = .1$	$n = 50$	$n = 100$	$n = 500$	$a_{21} = 1.0$	$n = 50$	$n = 100$	$n = 500$
$a_{22} = 0.8$	0.098	0.041	0.018	$a_{22} = 0.8$	0.091	0.124	0.013
$a_{22} = 0.9$	0.132	0.067	0.038	$a_{22} = 0.9$	0.142	0.166	0.029
$a_{22} = .95$	0.161	0.088	0.055	$a_{22} = .95$	0.206	0.099	0.038

Tables 2 and 3 show that the proportion of non-causal estimates are substantial, particularly for shorter series. In multiple settings over 10% of all datasets fitted resulted in non-causal estimates. Although we find other software options useful, utilizing our transformation provides

similar estimates but automatically guarantees all estimates to be causal. We also consider use of the MTS likelihood evaluation for gains in evaluation speed, although all reported results with our methodology use the exact VARMA likelihood.

## 6. DISCUSSION

In this article we introduced a new parameterization to describe the entire class of causal invertible VARMA processes in terms of unrestricted real-valued parameters. The proposed parameterization is as dense as the original VARMA parameterization in terms of nonzero parameters. The total number of parameters is the same as that of the original  $m$ -dimensional VARMA( $p, q$ ) process, i.e. equal to  $(p + q)m^2 + \binom{m+1}{2}$ . For moderately large  $p$  and  $q$ , or more realistically  $m$ , as with most multi-dimensional problems with dense parameterization, care needs to be exercised in computation. Sparse parameterizations that maintain causality and invertibility are a topic of future research. As with most transformation approaches, the transformed parameters lose interpretability. This is particularly true in the Bayesian setting where the prior is being invoked through the transformed parameters. However, unlike many other common constraint spaces, such as the positive definite matrix cone, there are no readily available distributions on the Schur-stable space, and truncation of known distributions, (e.g. multivariate normal on the original parameters) to the Schur-stable space lacks interpretability as well. The transformation method could be thought of as a tool for making inference under the desired restrictions. The final analysis of the VARMA estimation problem remains focused on the AR and MA polynomials in their original form.

The parameterization can be applied to many other estimation procedures as well, including quasi maximum likelihood methods, minimum distance methods and moment based methods. For example, an objective function measuring the closeness of the sample autocovariances and the theoretical autocovariances written in terms of the transformed parameters can be optimized to get causal invertible estimates. Moreover, one may also consider objective functions in the spectral domain, such as the integrated Frobenius norm of the difference between the observed periodograms and the spectral matrix. The advantage of the proposed parameterization is that any estimator, obtained as a minimizer of an objective function written in terms of transformed parameters, is

guaranteed to be causal and invertible. In future work we will demonstrate how estimators can be obtained with penalized likelihoods and our parameterization, thus simultaneously implementing sparsity constraints and causality constraints.

The proposed parameterization can potentially facilitate other aspects of VARMA modeling, such as a reduced rank formulation (Velu et.al.1986, Ahn and Reinsel 1988). The reduced rank version of the proposed parameterization is simple for a first order polynomial but for higher order polynomials the exact formulation needs to be investigated.

#### APPENDIX: PROOFS OF THEOREMS

*Proof of Theorem 1.* Suppose  $A(z)$  is Schur-Stable. Then

$$A(B^{-1})B^k X_t = Z_t$$

defines a Causal  $VAR(k)$  process. If we let  $\underline{U}_k = \underline{\Gamma}_k$  then  $A = \xi'_k \underline{U}_{k-1}^{-1}$  holds by the Yule-Walker equations for causal  $VAR(k)$  (see e.g., Brockwell and Davis, 1987, eq. 11.5.7, p 419) and we have the assertion of the if part of the theorem. Conversely, suppose there is a positive definite block Toeplitz matrix  $\underline{U}_k$  of the form (9) such that the coefficients of the polynomial  $A = [A_1, \dots, A_k]$  satisfy  $A = \xi'_k \underline{U}_{k-1}^{-1}$ . The determinantal equation  $\det(A(z)) = 0$  has the same roots as the eigenvalues of  $\tilde{A}$  ( $\tilde{A}$  is the *Companion matrix* of the monic polynomial  $A(z)$  and has the same characteristic polynomial as  $A(z)$ ; see Horn and Johnson (1985)) where  $\tilde{A}$  is defined in (16). We can see that  $\tilde{A}$  satisfies the system of equations

$$\underline{U}_{k-1} = \tilde{A} \underline{U}_{k-1} \tilde{A}' + \tilde{\Sigma} \tag{24}$$

where  $\tilde{\Sigma} = \begin{pmatrix} C_k & 0 \\ 0 & 0 \end{pmatrix}$ . Because  $\underline{U}_k$  is positive definite, so is  $C_k$ . We then obtain stability of  $\tilde{A}$ , modifying the argument of Stein (1952) slightly to show that as long as  $C_k$  is positive definite, the eigenvalues of  $\tilde{A}$  are strictly smaller than one in absolute value. Let  $w^* = (w_1^*, \dots, w_p^*)'$  be any left eigenvector of  $\tilde{A}$  corresponding to an eigenvalue  $\lambda$ . Here  $q^*$  denotes the complex conjugate transpose of the matrix  $q$ . First we show that  $w_1^* \neq 0$ . From the structure of  $\tilde{A}$  we have  $w_1^* A_{i-1} + w_i^* = \lambda w_{i-1}^*$  for  $1 < i < k$  and  $w_1^* A_k = \lambda w_k^*$ . Thus, if  $w_1^*$  is zero then the entire eigenvector is zero, leading to a

contradiction. Pre and post multiplying equation (24) by  $w^*$

$$(1 - |\lambda|^2)w^* \underline{U}_{k-1} w = w^* \tilde{\Sigma} w = w_1^* C_k w_1 > 0.$$

From the positive definiteness of  $\underline{U}_{k-1}$  it follows that  $|\lambda| < 1$ . Thus, since the eigenvalues of  $\tilde{A}$  are less than one in absolute value, we have the required Schur-stability of  $A(z)$ .  $\square$

*Proof of Theorem 2.* Let  $\underline{U}_k$  be a positive definite block Toeplitz matrix. Then  $C_k >_L 0$ . For  $2 \leq t \leq k$ ,

$$\begin{aligned} C_t &= U(0) - \xi_t' \underline{U}_{t-1}^{-1} \xi_t \\ &= U(0) - [\xi_{t-1}', U(t)'] \begin{pmatrix} \underline{U}_{t-2} & \kappa_{t-1} \\ \kappa_{t-1}' & U(0) \end{pmatrix}^{-1} \begin{bmatrix} \xi_{t-1} \\ U(t) \end{bmatrix} \\ &= U(0) - \xi_{t-1}' \underline{U}_{t-2}^{-1} \xi_{t-1} \\ &\quad - (U(t) - \xi_{t-1}' \underline{U}_{t-2}^{-1} \kappa_{t-1}) D_{t-1}^{-1} (U(t) - \xi_{t-1}' \underline{U}_{t-2}^{-1} \kappa_{t-1})' \\ &= C_{t-1} - (U(t) - \xi_{t-1}' \underline{U}_{t-2}^{-1} \kappa_{t-1}) D_{t-1}^{-1} (U(t) - \xi_{t-1}' \underline{U}_{t-2}^{-1} \kappa_{t-1})'. \end{aligned}$$

Here terms involving  $\xi_0, \kappa_0$  for the  $t = 1$  case are assumed to be zero. The second term on the right side is a nonnegative definite term. Since  $\underline{U}_k >_L 0$ , any principal diagonal block is positive definite and hence the Schur-complements  $D_t >_L 0$  for  $t \leq k$ . Thus,

$$C_{t-1} - C_t = (U(t) - \xi_{t-1}' \underline{U}_{t-2}^{-1} \kappa_{t-1}) D_{t-1}^{-1} (U(t) - \xi_{t-1}' \underline{U}_{t-2}^{-1} \kappa_{t-1})' \geq_L 0. \quad (25)$$

Based on the assumption that  $\underline{U}_k >_L 0$  we have  $C_k >_L 0$ . Hence  $C_0 \geq_L C_1 \geq_L \cdots \geq_L C_k >_L 0$ .

For the converse, let  $C_k >_L 0$ . Hence  $C_0 \geq_L C_1 \geq_L \cdots \geq_L C_k >_L 0$  be defined as in (15). Then  $C_0 = U(0) >_L 0$  and  $C_k = U(0) - \xi_k' \underline{U}_{k-1}^{-1} \xi_k >_L 0$  which implies

$$\underline{U}_k = \begin{pmatrix} U(0) & \xi_k' \\ \xi_k & \underline{U}_{k-1} \end{pmatrix},$$

is positive definite.  $\square$

*Proof of Theorem 3.* If  $A(z)$  is Schur-Stable construct a causal VAR( $k$ ) process  $\{X_t\}$  with innovation variance as the identity matrix  $I_m$  and  $A$  as the coefficient matrix. Then  $\underline{U}_{k-1} = \underline{\Gamma}_{k-1}$  is the required positive definite block Toeplitz matrix where  $\underline{\Gamma}_k$  is the variance  $\text{Var}\{X'_t, X'_{t-1}, \dots, X'_{t-k}\}$ .

For the converse suppose  $\underline{U}_{k-1} \in \mathfrak{T}_{++}^{m,k-1}$  exists with  $\tilde{S}_k(A, \underline{U}_{k-1}) \in \mathcal{S}_+^{mk}$  and  $S_k(A, \underline{U}_{k-1}) \in \mathcal{S}_{++}^m$ . Because of the block Toeplitz structure,

$$\tilde{S}_k(A, \underline{U}) = \begin{pmatrix} S_k(A, \underline{U}) & \star \\ \star & 0 \end{pmatrix}.$$

Positive definiteness of  $\tilde{S}_k(A, \underline{U}_{k-1})$  would then imply

$$\tilde{S}_k(A, \underline{U}) = \begin{pmatrix} S_k(A, \underline{U}) & 0 \\ 0 & 0 \end{pmatrix}.$$

Let  $w^* = (w_1^*, \dots, w_p^*)'$  be any left eigenvector of  $\tilde{A}$  corresponding to an eigenvalue  $\lambda$ . Because  $S_k(A, \underline{U}_{k-1}) \in \mathcal{S}_{++}^m$ , following the arguments given in the proof of Theorem 1, we have  $|\lambda| < 1$ .  $\square$

*Proof of Theorem 4.* The fact that the map  $\tau$  is a bijection is straightforward and follows from known properties of the Cholesky and Cayley transformations and the algorithm given in Gaillier (2013). We focus on showing that for each  $M$ , the maps  $\phi_M$  and  $\psi_M$  are bijections. Let  $\phi_M^{-1}$  be the inverse map of  $\phi_M$  where  $\phi^{-1}$  is defined by Theorem 1, i.e, for each  $\underline{U}_k \in \mathfrak{T}_{++}^{m,k}(M)$

$$\phi_M^{-1}(\underline{U}_k) = \xi'_k \underline{U}_{k-1}^{-1}.$$

Let  $\psi_M^{-1}$  denote the inverse map of  $\psi_M$  that is given by Algorithm [VQ]. We will show

$$\phi_M \circ \phi_M^{-1} = \phi_M^{-1} \circ \phi_M = id = \psi_M \circ \psi_M^{-1} = \psi_M^{-1} \circ \psi_M,$$

where  $id$  is the identity map of appropriate dimensions.

First consider any  $A \in \mathfrak{S}_k^m$ . By definition, the VAR( $k$ ) coefficients are the solutions to the Yule-Walker equation involving the variance matrix  $\underline{U}_k$ . Hence

$$A = (U(1)', \dots, U(k-1)') \underline{U}_{k-1}^{-1} = \phi^{-1}(\underline{U}_k) = \phi_M^{-1}(\phi_M(A)),$$

which implies  $\phi_M^{-1} \circ \phi_M = id$ . Next consider an arbitrary element  $\underline{U}_k$  of  $\mathfrak{S}_{++}^{m,k}(M)$ . Let  $A = \phi_M^{-1}(\underline{U}_k)$ .

Then  $\underline{U}_k$  satisfies the discrete algebraic Riccati equations

$$\underline{U}_k = \tilde{A}\underline{U}_k\tilde{A}' + \tilde{\Sigma},$$

where  $\tilde{A}$  is defined in (16) and  $\tilde{\Sigma} = \begin{pmatrix} M & 0 \\ 0 & 0 \end{pmatrix}$ . The Riccati equations are unique and the variance matrix of  $Y' = (Y'_1, \dots, Y'_k)$ , where  $Y_t$  is a VAR(k) process with coefficient matrices given by  $A$  and innovation variance  $M$ , also satisfies the Riccati equations. Therefore  $\phi_M \circ \phi_M^{-1} = id$ . Also for  $\underline{U}_k$  let

$$\psi_M(\underline{U}_k) = (V_1, \dots, V_k, Q_1, \dots, Q_k).$$

To apply the inverse map to  $\psi_M(\underline{U}_k)$  we apply Algorithm [VQ]. To distinguish the the quantities under the inverse map from the original quantities, we will use a  $\leftarrow$  accent on top of the inverse images. Since  $V_j$  are defined as the difference of the Schur complement sequence  $C_j$ , from the first step in the algorithm we have  $\overleftarrow{U}(0) = M + \sum_{j=1}^k V_j = C_k + \sum_{j=1}^k (C_{j-1} - C_j) = C_0 = U(0)$ . Since  $\overleftarrow{U}(j)$  are defined iteratively, we proceed by induction. Let  $\overleftarrow{U}(i) = U(i)$  for  $i = 0, 1, \dots, (j-1)$ . Then at stage  $j$ , we have

$$\begin{aligned} \overleftarrow{U}(j) &= \overleftarrow{\kappa}_{j-1} \overleftarrow{U}_{j-2} \overleftarrow{\xi}_{j-1} + V_j^{1/2} Q_j \overleftarrow{D}_{j-1}^{-1/2} \\ &= \overleftarrow{\kappa}'_{j-1} \overleftarrow{U}_{j-2}^{-1} \overleftarrow{\xi}_{j-1} + V_j^{1/2} Q_j \overleftarrow{D}_{j-1}^{1/2} \end{aligned}$$

Since  $V_j, Q_j$  are obtained under the map  $\psi_M$ , from (13) and (15) we have  $V_j^{1/2} Q_j \overleftarrow{D}_{j-1}^{1/2} = U(j)' - \overleftarrow{\kappa}'_{j-1} \overleftarrow{U}_{j-2}^{-1} \overleftarrow{\xi}_{j-1}$ . Thus  $\overleftarrow{U}_k = \underline{U}_k$  and hence  $\psi_M^{-1} \circ \psi_M = id$ . For the converse, let  $(V_1, \dots, V_k, Q_1, \dots, Q_k) \in (\mathcal{S}_{++}^m)^k \times O(m)^k$  be given. Then  $\underline{U}_k = \psi^{-1}(V_1, \dots, V_k, Q_1, \dots, Q_k)$  is constructed using Algorithm [VQ]. The inverse images  $(\overleftarrow{V}_1, \dots, \overleftarrow{V}_k, \overleftarrow{Q}_1, \dots, \overleftarrow{Q}_k)$  are defined using equations (13) and (15). Following similar arguments as before, it is straightforward to show by induction that the iterative scheme of defining  $\overleftarrow{V}_j, \overleftarrow{Q}_j$  yields  $(\overleftarrow{V}_j, \overleftarrow{Q}_j) = (V_j, Q_j)$ , for  $j = 1, \dots, k$ . Hence,  $\psi_M \circ \psi_M^{-1} = id$ .  $\square$

#### DISCLAIMER:

This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

## SUPPLEMENT DESCRIPTION

1. Appendix A: Simpler example using VAR(1)
2. Appendix B: Notes about implementation and additional simulation results

## REFERENCES

- [1] Athanasopoulos, G. and Vahid, F. (2008b), “VARMA versus VAR for macroeconomic forecasting”, *Journal of Business and Economic Statistics* **26**(2), pp. 237 – 252.
- [2] Ahn, S. K. and Reinsel, G. C. (1988), “Nested Reduced Rank Autoregressive Models for Multiple Time Series,” *Journal of the American Statistical Association*, 83, 849–856.
- [3] Albert, J. (2009), *Bayesian Computation with R*. Springer, New York, second edition.
- [4] Ansley, C. F. (1988), “An Algorithm for the Exact Likelihood of Mixed Autoregressive Moving Average Process,” *Biometrika*, 66, 59–65.
- [5] BARNDORFF-NIELSEN, O. and SCHOU, G. (1973), “On the Parametrization of Autoregressive Models by Partial Autocorrelations.” *Journal of Multivariate Analysis* **3** 408–419.
- [6] Bhatia, R. (1997), *Matrix Analysis*, Springer.
- [7] Breidt, F. J., Davis, R. A., Lii, K. S. and Rosenbalatt, M. (1991). “Maximum likelihood estimation for noncausal autoregressive processes”. *Journal of Multivariate Analysis*, 36 175–198.
- [8] Breidt, F. J., Davis, R. A. and Trindade, A. A. (2001). “Least absolute deviation estimation for all-pass time series models”. *The Annals of Statistics*, 29 919–946.
- [9] Brockwell, P.J. and Davis, R.A. (1991), *Time Series: Theory and Methods*, 2nd ed. Springer.
- [10] Chikuse, Y. (2003), *Statistics on Special Manifolds, Volume 174 of Lecture Notes in Statistics*, Springer-Verlag, New York.

- [11] Chiuso, A., Giorgio, P. and Soatto, S. (2008), “Wide-sense Estimation on the Special Orthogonal group”, *Communications in Information and Systems*, 8, 185–200.
- [12] Constantinescu, T. (1986), “Schur Analysis of Positive Block-Matrices”, In *I. Schur methods in operator theory and signal processing*, volume 18 of *Oper. Theory Adv. Appl.*, 191–206.
- [13] Chung, Y., Rabe-Hesketh, S., Gelman, A., Liu, J. and Dorie, V. (2011), “Avoiding Boundary Estimates in Linear Mixed Models Through Weakly Informative Priors”, *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 284.
- [14] Delsarte, P., Genin, Y. and Kamp, Y. (1979), “Schur Parametrization of Positive Definite Block-Toeplitz Systems”, *SIAM Journal of Applied Math*, 36, 34–46.
- [15] Doan, T., Litterman, R. and Sims, C. (1984), “Forecasting and Conditional Projection Using Realistic Prior Distributions”, *Econometric Reviews*, 3, 1–100.
- [16] Dunson, D.B. and Neelon, B. (2003). “Bayesian Inference on Order-Constrained Parameters in Generalized Linear Models”, *Biometrics*, 59, 286–295.
- [17] Fuller, W. A. (1995), *Introduction to Statistical Time Series*, Wiley-Interscience.
- [18] Gallier, J. (2013), “Remarks on the Cayley Representation of Orthogonal Matrices and on Perturbing the Diagonal of a Matrix to Make it Invertible”, [online], Available at: [arXiv:math/0606320](https://arxiv.org/abs/math/0606320).
- [19] Gelfand, A. E., Smith, A. F. M. and Lee, T. M. (1992), “Bayesian Analysis of Constrained Parameters and Truncated Data Problems”, *Journal of the American Statistical Association*, 87, 523–532.
- [20] Giurcanu, M. (2015). A simulation algorithm for non-causal VARMA processes. *Statistics & Probability Letters*, 98 65–72.
- [21] Godolphin, E. J. (1984), “A Direct Representation for the Large Sample Maximum Likelihood Estimator of a Gaussian Autoregressive Moving Average Process”, *Biometrika*, 71, 281–89.

- [22] Gouriéroux, C. and Jasiak, J. (2016). Filtering, Prediction and Simulation Methods for Non-causal Processes. *Journal of Time Series Analysis*, 37 405–430.
- [23] Gouriéroux, C. and Zakoan, J.-M. (2016). Local explosion modelling by non-causal process. *Journal of the Royal Statistical Society: Series B*. doi: 10.1111/rssb.12193
- [24] Hannan, E. J. and Rissanen J. (1982), “Recursive Estimation of Mixed Autoregressive-Moving Average Order”, *Biometrika*, 69, 81–94.
- [25] Hoff, P.D. (2009), “Simulation of the Matrix Bingham-von Mises-Fisher Distribution, With Applications to Multivariate and Relational Data”. *Journal of Computational and Graphical Statistics*, 18, 438–456.
- [26] Kadiyala, K. R. and Karlsson, S. (1997), “Numerical Methods for Estimation and Inference in Bayesian VAR-Models”, *Journal of Applied Econometrics*, 12, 99–132.
- [27] Kaszkurewicz, E. and Bhaya, A. (2000), *Matrix Diagonal Stability in Systems and Computation*, Birkhauser, Boston.
- [28] Koop, G. and Korobilis, D. (2009), *Bayesian Multivariate Time-Series Methods in Empirical Macroeconomics*. Now Publishers Inc, Boston.
- [29] Koop, G. and Potter S. M. (2011), “Time Varying VARs with Inequality Restrictions”, *Journal of Economic Dynamics and Control*, 35, 1126–1138.
- [30] Koreisha, S. and Pukkila, T. (1989), “Fast Linear Estimation Methods for Vector Moving Average Models”, *Journal of Time Series Analysis*, 10, 325–339.
- [31] Lanne, M. and Saikkonen, P. (2013). Noncausal vector autoregression. *Econometric Theory*, 29 447–481.
- [32] Leonard, T. and Hsu, J. S. J. (2002), “Bayesian Inference for a Covariance Matrix”, *Annals of Statistics*, 20, 1669–1696.

- [33] Lindstrom, M. J. and Bates, D. M. (1988), “Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data”, *Journal of the American Statistical Association*, 83, 1014–1022.
- [34] Litterman, R. B. (1980), “Techniques for Forecasting with Vector Autoregressions”, *Ph.D. thesis, University of Minnesota*.
- [35] Marin, J.-M. and Robert, C. P. (2007), *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer, New York.
- [36] Marriott, J.M. and Smith, A.F.M. (1992), “Reparametrization Aspects of Numerical Bayesian Methodology for Autoregressive Moving-Average Models”, *Journal of Time Series Analysis*, 13, 327–343.
- [37] Mauricio, J. A. (1995). “Exact Maximum Likelihood Estimation of Stationary Vector ARMA Models”, *Journal of the American Statistical Association*, 90, 282–291.
- [38] Mauricio, J. A. (1997), “The Exact Likelihood of a Vector Autoregressive Moving Average Model”, *Applied Statistics*, 46, 157–171.
- [39] Mauricio, J. A. (2002), “An Algorithm for the Exact Likelihood of a Vector Autoregressive Moving Average Model”, *Journal of Time Series Analysis*, 23, 473–486.
- [40] Metaxoglou, K. and Smith, A. (2007), “Maximum Likelihood Estimation of VARMA Models Using a State-Space EM Algorithm”, *Journal of Time Series Analysis*, 28, 666–685.
- [41] Nelson, C. R. and Plosser, C. I. (1982), “Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications”, *Journal of Monetary Economics*, 10, 139–162.
- [42] Nicholls, D. F. and Hall, A. D. (1979), “The Exact Likelihood Function of Multivariate Autoregressive Moving Average Models”, *Biometrika*, 66, 259–264.
- [43] Nyberg, H., Lanne, M. and Saarinen, E. (2012). “Does noncausality help in forecasting economic time series?” *Economics Bulletin*, 32, 2849–2859.

- [44] Nyberg, H. and Saikkonen, P. (2014) “Forecasting with noncausal VAR model”. *Computational Statistics & Data Analysis*, 76 C, 536–555.
- [45] Pinheiro, J. C. and Bates, D. M. (1996), “Unconstrained Parametrizations for Variance-Covariance Matrices”, *Statistics and Computing*, 6, 289–296.
- [46] Quenneville, B. and McLeod, A. I. (1992), “Integration Over the Stationary and Invertible Region of an Autoregressive Moving Average Process”, *Proceedings of the Joint Statistical Meetings, Boston, Massachusetts*.
- [47] Reinsel, G. C., Basu, S. and Yap, S. F. (1992), “Maximum Likelihood Estimators in the Multivariate Autoregressive Moving Average Model from a Generalized Least Squares Viewpoint”, *Journal of Time Series Analysis*, 13, 133–145.
- [48] Schneider, H. (1965), “Positive Operators and an Inertial Theorem”, *Numer. math.*, 7, 11–17.
- [49] Simionescu, M. “The Use of VARMA Models in Forecasting Macroeconomic Indicators,” *Economics & Sociology*, Vol. 6, No 2, 2013, pp. 94- 102. DOI: 10.14254/2071-789X.2013/6-2/9
- [50] Sims, C. A. and Zha, T. (1998), “Bayesian Methods for Dynamic Multivariate Models,” *International Economic Review*, 39, 949–968.
- [51] Sims, C. A. (2010), VAR Tools webpage. <http://sims.princeton.edu/yftp/VARtools/>
- [52] Stein, P. (1952), “Some General Theorems on Iterants”, *J. Res. Natl. Bur. Standards*, 48, 82-83.
- [53] Tunnicliffe-Wilson, G. (1973), “The Estimation of Parameters in Multivariate Time Series Models”, *Journal of the Royal Statistical Society, B*, 35, 76-85.
- [54] Tsay, R. (2014) *Multivariate Time Series Analysis With R and Financial Applications*, Wiley.
- [55] Tsay, R. (2015), MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models. R package version 0.33. <http://CRAN.R-project.org/package=MTS>

- [56] Velu, R. P., Reinsel, G. C. and Wichern, D. W. (1986), “Reduced Rank Models for Multiple Time Series”, *Biometrika*, 73, 105–118.
- [57] Whittle, P. (1951), *Hypothesis Testing in Time Series Analysis*. Upsala: Almqvist and Wiksell.
- [58] Wise, J. (1956), “Stationarity Conditions for Stochastic Processes of the Autoregressive and Moving-Average Type”, *Biometrika*, 43, 215–219.
- [59] Zadrozny, P. (1998) An eigenvalue method of undetermined coefficients for solving linear rational expectations models. *Journal of Economic Dynamics and Control* 22, 1353–1373.