Statistica Sinica **29** (2019), 455-478
doi:https://doi.org/10.5705/ss.202016.0415

CONSTRAINED ESTIMATION OF CAUSAL INVERTIBLE VARMA

Anindya Roy[1], Tucker S. Mcelroy[2] and Peter Linton[1]

[1] *University of Maryland Baltimore County and* [2] *U.S. Census Bureau*

Abstract: 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.

Key words and phrases: Block Toeplitz matrix, constrained estimation, reparameterization, Schur stability.

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 such special cases as the Yule-Walker for vector autoregressions). We fill 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), Gourieroux and Jasiak (2016)) and non-Gaussian models (Breidt et al. (1991); Breidt, Davis and Trindade (2001), Lanne and Saikkonen (2013); Giurcanu (2015); Nyberg, Lanne and Saarinen (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

$$\Phi(B)X_t = \Theta(B)Z_t, \tag{1.1}$$

indexed by time t. The innovations, $\{Z_t\}$, satisfy $\mathrm{E}(Z_t) = 0$, $\mathrm{Var}(Z_t) = \Sigma$, and $\mathrm{Cov}(Z_s, Z_t) = 0$ 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

$$\Phi(z) = I_m - \Phi_1 z - \cdots - \Phi_p z^p, \tag{1.2}$$

$$\Theta(z) = I_m + \Theta_1 z + \cdots + \Theta_q z^q, \tag{1.3}$$

with $m \times m$ coefficient matrices $\Phi_1, \ldots, \Phi_p$ and $\Theta_1, \ldots, \Theta_q$, 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 Γ such that $\Gamma(k) = \mathrm{E}(X_t X'_{t-k})$ 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 (1.1) 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 (1.1), 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 unity; 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 (1995, 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 complexity 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-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, Basu and Yap (1992)). Other authors (Mauricio (1995, 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 to Bayesian vector autoregression (BVAR) and prior specification for such models. The popular choices for prior specification include those described in Litterman (1980), Doan, Litterman and Sims (1984), Kadiyala and Karlsson (1997), and Sims and Zha (1998). These formulations usually have normal priors on the coefficients along with inverse Wishart priors 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, as 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 in BVAR; (Wise (1956)). 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-Nielsen 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 mini-

mum 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 Section S2 of the supplementary material. Additional numerical illustrations as well as notes on computational aspects of the proposed algorithm are given in Section S3 in the supplement.

2 Schur-stability

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, Thm. 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

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

A VARMA process defined by (1.1) 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 (1.1) 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, let

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

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 Σ as the parameters of the process defined by (1.1) and when it is clear we interchangeably refer to the associated coefficient matrices $\Phi = (\Phi_1, \ldots, \Phi_p)$, $\Theta = (\Theta_1, \ldots, \Theta_q)$ and Σ 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. Take

$$\mathscr{S}^m_{++} = \{\Sigma \in \mathscr{S}^m : \Sigma >_L 0\} \tag{2.1}$$

to be the set of all $m \times m$ symmetric positive definite matrices that constitute the interior of the convex cone, $\mathscr{S}^m_+$, of $m \times m$ positive semi-definite matrices in $\mathscr{S}^m$, the set of $m \times m$ symmetric matrices. A matrix monic polynomial $A(z) = z^k I_m - A_1 z^{k-1} - \cdots - A_k$, is 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

$$\begin{aligned}\mathfrak{S}^{m,k} &= \{A(z) = z^k I_m - A_1 z^{k-1} - \cdots - A_k : A_r \in \mathbb{R}^{m\times m}, \\ &\qquad r \geq 1, \text{and } A(z) \text{ is Schur-stable}\} \end{aligned} \tag{2.2}$$

define the set of all $m-$dimensional Schur-stable matrix monic polynomials of degree k, and let any polynomial $A(z) = z^k I_m - A_1 z^{k-1} - \cdots - A_k$ be associated with the coefficient sequence $A = [A_1, \ldots, A_k]$. We take a sequence of matrices $A = [A_1, \ldots, A_k]$ 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

$$\mathfrak{P} = \mathfrak{S}^{m,p} \times \mathfrak{S}^{m,q} \times \mathscr{S}^m_{++}. \tag{2.3}$$

Often for a VARMA(p, q) process the innovations are assumed to be Gaussian, $Z_t \sim N(0, \Sigma)$. Based on this, 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 $p \geq 0$, let $\underline{\Gamma}_p$ be the covariance matrix of $(X_t', X_{t-1}', \ldots, X_{t-p}')'$. The jkth block of $\underline{\Gamma}_p$ is an $m \times m$ matrix, given by $\Gamma(k-j) = \Gamma'(j-k)$, for $1 \leq j, k \leq (p+1)$. If $Z_t \overset{i.i.d.}{\sim} N(0, \Sigma)$, a stationary likelihood for (Φ, Θ, Σ) based on a sample $X = (X_n', \ldots, X_1')'$ (written in the reverse order for notational consistency) is

$$\mathcal{L}(\Phi, \Theta, \Sigma) = (2\pi)^{-n/2} \left\{\det(\underline{\Gamma}_{n-1})\right\}^{-1/2} \exp\left(-0.5 X' \underline{\Gamma}_{n-1}^{-1} X\right). \tag{2.4}$$

where $\underline{\Gamma}_{n-1}$ is a function of (Φ, Θ, Σ). An available likelihood, evaluated through $\underline{\Gamma}_{n-1}$, immediately facilitates estimation. For causal invertible processes, maximum likelihood estimators of (Φ, Θ, Σ) 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 (Φ, Θ, Σ).

The Schur-stable space $\mathfrak{S}^{m,k}$ is described by the roots of the matrix polynomials. Unfortunately, the roots are highly non-linear functions of parameters Φ and Θ, and often are implicitly defined. Thus, direct maximization of the likelihood (2.4) 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, Smith and Lee (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

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

For $j \geq 1$, take $\underline{U}_j$ to be a symmetric block Toeplitz matrix of order j:

$$\underline{U}_j = \begin{pmatrix} U(0) & U(1) & \cdots & U(j) \\ U(1)' & U(0) & \cdots & \cdot \\ \vdots & \ddots & \ddots & \vdots \\ U(j)' & \cdots & U(1)' & U(0) \end{pmatrix}, \tag{3.1}$$

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

$$\underline{U}_j = \begin{pmatrix} U(0) & \xi_j' \\ \xi_j & \underline{U}_{j-1} \end{pmatrix}, \tag{3.2}$$

and the upper representation given by

$$\underline{U}_j = \begin{pmatrix} \underline{U}_{j-1} & \kappa_j \\ \kappa_j' & U(0) \end{pmatrix}. \tag{3.3}$$

Here $\xi_j' = (U(1), \ldots, U(j))$ and $\kappa_j' = (U(j)', \ldots, U(1)')$. Set ξ_0 and κ_0 equal to zero matrices. The Schur complements of $\underline{U}_{j-1}$ in $\underline{U}_j$ in the two representations (3.2) and (3.3) are

$$C_j = U(0) - \xi_j' \underline{U}_{j-1}^{-1} \xi_j, \tag{3.4}$$

$$D_j = U(0) - \kappa_j' \underline{U}_{j-1}^{-1} \kappa_j. \tag{3.5}$$

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

$$\mathfrak{T}^{m,k} = \{\underline{U}_k \in \mathscr{S}^{m(k+1)} : \underline{U}_k \text{ is in the form (3.1)}\}$$

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 $A(z) = I_m z^k - A_1 z^{k-1} \cdots - A_k$ is Schur-stable if and only if there exists $\underline{U}_k \in \mathfrak{T}_{++}^{m,k}$ such that the coefficients $A = [A_1, \ldots, A_k] \in \mathbb{R}^{m \times mk}$ 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) and Delsarte, Genin and Kamp (1979). We can describe 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. This characterization allows the successive difference of the Schur complements, $C_{i-1} - C_i$ to be nonnegative definite. For implementation we 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 $1 \le j \le k$. Following the proof of Theorem 2,

$$V_j = \left\{U(j) - \xi'_{j-1}\underline{U}_{j-2}^{-1}\kappa_{j-1}\right\}D_{j-1}^{-1}\left\{U(j) - \xi'_{j-1}\underline{U}_{j-2}^{-1}\,\kappa_{j-1}\right\}', \tag{3.6}$$

which has the solution

$$U(j) = \xi'_{j-1}\underline{U}_{j-2}^{-1}\kappa_{j-1} + V_j^{1/2}Q_jD_{j-1}^{1/2} \tag{3.7}$$

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

$$Q_j = V_j^{-1/2}\left\{U(j) - \xi'_{j-1}\underline{U}_{j-2}^{-1}\kappa_{j-1}\right\}D_{j-1}^{-1/2}. \tag{3.8}$$

Here $V_j^{1/2}$ and $D_{j-1}^{1/2}$ are square roots of V_j and D_{j-1}, respectively. Then (3.7) defines the key recursion equation that allows one to solve for $U(j)$, $j = 0, \ldots, k$ iteratively, once the positive definite matrices $M, V_1, \ldots, V_k$ and the orthogonal matrices $Q_1, \ldots, Q_k$ have been specified. At the jth stage, all quantities on the right side of (3.7) are known; thus $U(j)$ and $\underline{U}_j$ can be computed. Subsequently ξ_j, κ_j and D_j are obtained from $\underline{U}_j$ through (3.4) and (3.5), and then used in the $(j+1)$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 $V_1, \ldots, V_k, Q_1, \ldots, Q_k$

1. Set $U(0) = C_0 = M + \sum_{j=1}^{k} V_j$.
2. Compute $U(1)' = V_1^{1/2}\,Q_1U(0)^{1/2}$ and obtain $\underline{U}_1$.
3. Compute κ_1, ξ_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 $U(2) = \xi'_1U(0)^{-1}\kappa_1 + V_2^{1/2}Q_2D_1^{1/2}$ and obtain $\underline{U}_2$ from $U(0)$, $U(1), U(2)$.
5. Obtain κ_2, ξ_2, D_2 and iterate using (3.7).
6. Once $U(0), U(1), \ldots, U(k)$ and hence $\underline{U}_k$ have been obtained, compute A using the Yule-Walker relation $A = \xi'_k\underline{U}_{k-1}^{-1}$.

The inverse algorithm is obtained by noting that $\mathrm{Vec}(\underline{U}_{k-1}) = (I - \tilde{A} \otimes$

$\tilde{A})^{-1}\text{Vec}(\tilde{M})$ and $\xi_k' = A\underline{U}_{k-1}$ where I is the identity of dimension $(mk \times mk)$,

$$\tilde{A} = \begin{pmatrix} A_1 & A_2 & \cdots & A_{k-1} & A_k \\ I_m & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & I_m & 0 \end{pmatrix}, \tag{3.9}$$

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

3.2. Generalized Stein transformation

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

$$S(A, U) = U - AUA'. \tag{3.10}$$

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 $\mathscr{S}^m \to \mathscr{S}^m$. 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 $U \in \mathscr{S}^m_{++}$ such that $S(A, U) \in \mathscr{S}^m_{++}$ if and only if $A \in \mathfrak{S}^m_1$. Stein's result implies that one could characterize $\mathfrak{S}^{m,1}$ in terms of matrices in $\mathscr{S}^m_{++}$. For any $M \in \mathscr{S}^m_{++}$, 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 $A = [A_1, \ldots, A_k] \in \mathbb{R}^{m\times mk}$ associated with a polynomial $A(z) = I_m z^k - A_1 z^{k-1} - \cdots - A_k$, and a symmetric matrix $\underline{U} \in \mathscr{S}^{mk}$ with U_{11} as the upper left $m \times m$ block of $\underline{U}$. Define the *Generalized Stein Transformation* $\tilde{S}_k(A, \underline{U}) : \mathbb{R}^{m\times mk} \times \mathscr{S}^{mk} \to \mathscr{S}^{mk}$ by

$$\tilde{S}_k(A, \underline{U}) = \underline{U} - \tilde{A}\underline{U}\tilde{A}', \tag{3.11}$$

so $\tilde{S}_k(A, \underline{U}) = S(\tilde{A}, \underline{U})$. Take $S_k(A, \underline{U}) = U_{11} - A\underline{U}A'$ 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 $A(z) = z^k - A_1 z^{k-1} - \cdots - A_k$ with coefficients $A = [A_1, \ldots, A_k]$ 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 $\tilde{S}_k(A, \underline{U}_{k-1}) \in \mathscr{S}^{mk}_+$ and $S_k(A, \underline{U}_{k-1}) \in \mathscr{S}^m_{++}$.*

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 $M \in \mathscr{S}^m_{++}$, 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 $A = \xi'_k \underline{U}^{-1}_{k-1}$ will be a member of the pre-image. Additionally,

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

The result is established by noting that given $A \in \mathfrak{S}^{m,k}$ and $M \in \mathscr{S}^m_{++}$, 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, Σ) 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 use-

ful. The specific form for the $m \times m$ positive definite matrix V is

$$V = LDL', \tag{3.12}$$

where L and D are given by

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

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 $l = (l_{2,1}, \ldots, l_{m,m-1})$ and m unconstrained real numbers in $d = (d_1, \ldots, d_m)$, 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 $e_1 = (1, 0, \ldots, 0)'$. 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. Gallier (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*

$$Q = E_\delta[(I_m - S)(I_m + S)^{-1}]^2, \tag{3.13}$$

for some skew-symmetric matrix S and some reflection E_δ. The inverse transformation that determines S from R can be defined in an injective manner following Proposition 1.3 in Gallier (2013). Let $S = 2(I_m + R^{1/2})^{-1} - I_m$ be the inverse image of R for the transformation (3.13), 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 Gallier (2013). The mapping of Q to the $(m(m-1))/2$ unrestricted elements in $s = (s_{21}, \ldots, s_{m,m-1})$ and the binary parameter δ is a bijection. Thus, if we define the map

$$\tau(V, Q) = (l, d, s, \delta) \tag{3.14}$$

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

3.5. Family of bijections

The Cholesky transformation (3.12) and the transformation (3.13) provide the final steps in mapping the coeffcient 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 $m \times m$ postive definite matrix M, there exists a bijection $f_M : \mathfrak{S}^{m,k} \mapsto \mathbb{R}^{km^2} \times \{0,1\}^k$. The bijection can be decomposed as $f_M = \tau \circ \psi_M \circ \phi_M$ where*

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

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

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

All are bijections and can be described as follows.

1. *For each $A \in \mathfrak{S}^{m,k}$, $\phi_M(A)$ is $\underline{U}_k$, the variance matrix of $Y' = (Y_1', \ldots, Y_k')'$ where $\{Y_t\}$ is a VAR(k) process with coefficient matrices given by A and innovation variance equal to M.*

2. *For each $\underline{U}_k \in \mathfrak{T}_{++}^{m,k}(M)$, $\psi_M(\underline{U}_k) = (V_1, \ldots, V_k, Q_1, \ldots, Q_k)$ where V_j, Q_j, $j = 1, \ldots, k$, are given iteratively at* (3.6) *and* (3.8).

3. *For each set of $(V_1, \ldots, V_k, Q_1, \ldots, Q_k) \in (\mathscr{S}_{++}^m)^k \times O(m)^k$ the map τ is*

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

The quantities $\underline{l}$ and $\underline{d}$ are the parameters of the Cholesky transformation (3.12) *for the set of all V_j and the quantities $\underline{s}$ and $\underline{\delta}$ are the parameters of the transformation* (3.13) *for the set of all Q_j.*

4. Numerical Computation

Numerical Computation and Estimation

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 are denoted as $(V_1^1,\ldots,V_p^1)$ and $(Q_1^1,\ldots,Q_p^1)$, whereas those associated with the moving average part are denoted as $(V_1^2,\ldots,V_q^2)$ and $(Q_1^2,\ldots,Q_q^2)$, respectively. The innovation variance is denoted by Σ. The matrix M, needed to start the parameterization of the Schur-stable polynomials is 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 are denoted as $\underline{l}^{1,j}=(l_{2,1}^{1,j},\ldots,l_{m,m-1}^{1,j})$ and $\underline{d}^{1,j}=(d_1^{1,j},\ldots,d_m^{1,j})$ for $j=1,\ldots,p$, and as $\underline{l}^{2,j}=(l_{2,1}^{2,j},\ldots,l_{m,m-1}^{2,j})$ and $\underline{d}^{2,j}=(d_1^{2,j},\ldots,d_m^{2,j})$, $j=1,\ldots,q$ for the moving average part. Similarly the scalar parameters associated with the orthogonal matrices are denoted as $\underline{s}^{1,j}=(s_{2,1}^{1,j},\ldots,s_{m,m-1}^{1,j})$ and $\underline{\delta}^{1,j}$, $j=1,\ldots,p$, for the autoregressive part, and as $\underline{s}^{2,j}=(s_{2,1}^{2,j},\ldots,s_{m,m-1}^{2,j})$ and $\underline{\delta}^{2,j}$ for $j=1,\ldots,q$ for the moving average part. The Cholesky parameters associated with the innovation variance are written as $\underline{l}^0=(l_{2,1}^0,\ldots,l_{m,m-1}^0)$ and $\underline{d}^0=(d_1^0,\ldots,d_m^0)$. All together the real scalar parameters are then $(\underline{l},\underline{d},\underline{s})$ where $\underline{l}=(\underline{l}^{1,1},\ldots,\underline{l}^{1,p},\underline{l}^{2,1},\ldots,\underline{l}^{2,q},\underline{l}^0)$, 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 δ parameters $\underline{\delta}=(\underline{\delta}^{1,1},\ldots,\underline{\delta}^{1,p},\underline{\delta}^{2,1},\ldots,\underline{\delta}^{2,q})$ lie in $\{0,1\}^{p+q}$ providing 2^{p+q} possible values for the parameters. Suppose the VARMA likelihood is written as $\mathcal{L}(\Phi,\Theta,\Sigma)$, 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

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

and hence that of the matrix parameters are

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

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 δ 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 Φ and Θ, 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 δ 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 is 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, Giorgio and Soatto (2008)).

5. Simulation and Data Analysis

Simulation and Data Analysis 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:

$$X_t = \begin{pmatrix} \Phi_{11} & 0 \\ 1 & 0.8 \end{pmatrix} X_{t-1} + Z_t, \tag{5.1}$$

where $Z_t \overset{iid}{\sim} N(0, I_2)$. We varied the upper diagonal entry Φ_{11}, because it is also one of the eigenvalues of Φ_1. The values of Φ_{11} were chosen from the set $\{-0.95, -0.95, -0.9, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 0.9,\ 0.95, 0.99\}$ to examine performance over a varying parameter space. Maximum likelihood estimation was done using the *optim* L-BFGS-B function in R. The initial values of the pre-parameters were chosen to be those calculated from the Yule-Walker estimate. We used box constraints with upper and lower bounds for the real parameters. The bounds were $\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 would prevent the iterations from generating near singular values of the V matrix. For Bayesian estimation, the $N(0, 5)$ prior was specified for all the pre-parameters except δ, which was assigned a $Bernoulli(0.5)$ prior. The Bayesian updates were done with Metropolis random walk for the real parameters and via independent sampling with a jump distribution of $Bernoulli(0.5)$ for δ. Metropolis chains reported were for length 20,000, with a burn-in of 5,000.

Let the Mean Square Error for the VAR(p) polynomial coefficient matrices be defined as $N^{-1} \sum_{j=1}^{N} \sum_{k=1}^{p} \|\hat{\Phi}_{j,k} - \Phi_k\|^2$, where $\|\cdot\|$ is the Frobenius norm of a matrix and $\hat{\Phi}_{j,k}$ is the jth Monte Carlo estimate of Φ_k. For a single entry in a coefficient matrix the Monte Carlo MSE was defined as the average

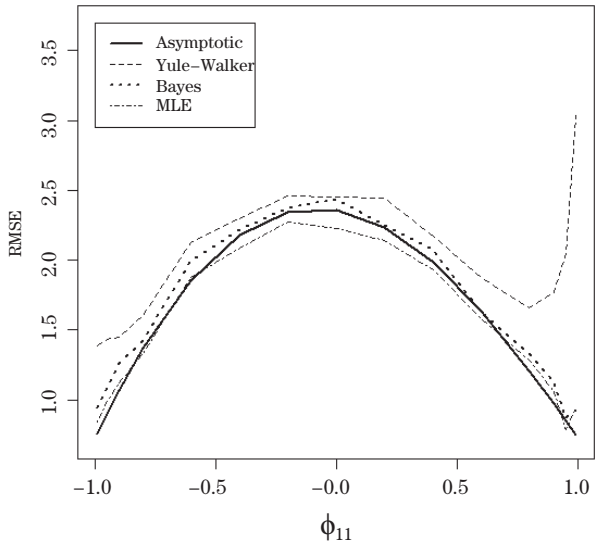

Figure 1. Overall RMSE, $nN^{-1}\sum_{j=1}^{N}\|\hat{\Phi}-\Phi\|^2$, for different estimators of Φ compared with the corresponding asymptotic value. The RMSE is plotted as a function of Φ_{11} in model (5.1).

square distance of the estimated value to the true value where the average is over $N = 1{,}000$ Monte Carlo replications. We report the square root of the MSE (RMSE) as a function of Φ_{11} and compare the overall RMSE with the asymptotic value obtained based on the asymptotic variance of the individual entries of Φ. The RMSE is plotted as a function of Φ_{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),

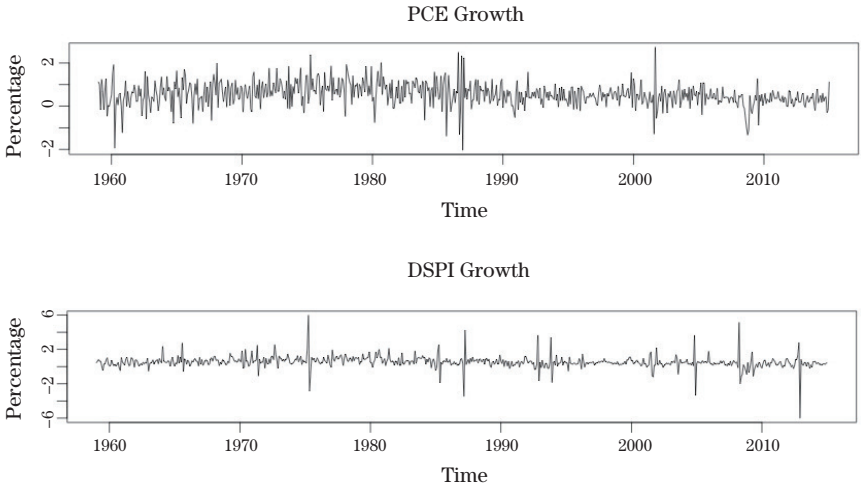

Figure 2. Plots of log differenced monthly personal consumption expenditure (PCE) and disposable personal income (DSPI).

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	0.445	0.272	0.356	0.058	0.135	0.098	−0.062
	(0.093)	(0.067)	(0.130)	(0.115)	(0.050)	(0.040)	(0.067)	(0.056)
MLE	0.457	0.411	0.284	0.357	0.060	0.139	0.116	−0.037
	(0.060)	(0.049)	(0.104)	(0.092)	(0.037)	(0.026)	(0.052)	(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	0.191	0.091	−0.070	−0.675	−0.345	−0.195	−0.642
	(0.045)	(0.044)	(0.065)	(0.059)	(0.084)	(0.060)	(0.121)	(0.110)
MLE	−0.012	0.186	0.096	−0.069	−0.650	−0.311	−0.199	−0.642
	(0.030)	(0.030)	(0.046)	(0.044)	(0.054)	(0.043)	(0.092)	(0.086)

a VARMA(3,1) model with Gaussian innovations fits the log differenced series reasonably well. We modeled the differenced log-scale series multiplied by 100 and the growth rates in percentages. 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).

The two methods yielded similar estimates with comparable standard errors, with the MLE generally having smaller standard errors. After adjusting for multiple comparison, the list of coefficients that were deemed statistically signif-

icantly different from zero was 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

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

for the two methods. For the proposed method, the standard errors of the estimated parameters in the reparametrized form were 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 were obtained by using a numerical linearization of the transformation and applying the delta method.

5.3. 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 Φ polynomial frequently estimate non-causal models. To demonstrate this phenomenon we simulated 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, 1,000 series of length n were generated with an autoregressive parameter using a local-to-unity parameterization, with the largest root and first element of Φ equal to $(1-1/n)$ for data of length n. The data generating process had Gaussian error and was 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 where fits were non causal for various n, a_{21} and a_{22} are given in Table 2.

One thousand datasets were 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 (5.2). The moving average parameter was fixed for all cases. The error was Gaussian. The proportion of cases for which the estimate was non-causal is given in Table 3.

$$\Phi = \begin{pmatrix} (1-1/n) & 0 \\ a_{21} & a_{22} \end{pmatrix} \quad \Theta = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}. \tag{5.2}$$

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

$a_{21} = 0.1$	$n = 50$	$n = 100$	$n = 500$
$a_{22} = 0.8$	0.193	0.096	0.022
$a_{22} = 0.9$	0.298	0.127	0.025
$a_{22} = 0.95$	0.367	0.186	0.039
$a_{21} = 1.0$	$n = 50$	$n = 100$	$n = 500$
$a_{22} = 0.8$	0.197	0.090	0.020
$a_{22} = 0.9$	0.262	0.128	0.040
$a_{22} = 0.95$	0.354	0.193	0.051

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

$a_{21} = 0.1$	$n = 50$	$n = 100$	$n = 500$
$a_{22} = 0.8$	0.098	0.041	0.018
$a_{22} = 0.9$	0.132	0.067	0.038
$a_{22} = 0.95$	0.161	0.088	0.055
$a_{21} = 1.0$	$n = 50$	$n = 100$	$n = 500$
$a_{22} = 0.8$	0.091	0.124	0.013
$a_{22} = 0.9$	0.142	0.166	0.029
$a_{22} = 0.95$	0.206	0.099	0.038

Tables 2 and 3 show that the proportion of non-causal estimates is 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, $(p + q)m^2 + \binom{m+1}{2}$. 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 transofrmation 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, Reinsel and Wichern (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.

Supplementary Materials

1. S1: Proofs of Theorems
2. Simpler example using VAR(1)
3. S3: Notes about implementation and additional simulation results

Acknowledgements and Disclaimer

The authors are grateful to the anonymous referees and the editors for their

helpful comments and suggestions. 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.

References

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

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.

Albert, J. (2009). *Bayesian Computation with R*. 2nd Edition, Springer, New York.

Ansley, C. F. (1988). An algorithm for the exact likelihood of mixed autoregressive moving average process. *Biometrika* **66**, 59–65.

Barndorff-Nielsen, O. and Schou, G. (1973). On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis* **3**, 408–419.

Bhatia, R. (1997). *Matrix Analysis*. Springer.

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.

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.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. 2nd Edition, Springer.

Chikuse, Y. (2003). *Statistics on Special Manifolds*. Volume 174 of Lecture Notes in Statistics, Springer-Verlag, New York.

Chiuso, A., Giorgio, P. and Soatto, S. (2008). Wide-sense estimation on the special orthogonal group. *Communications in Information and Systems* **8**, 185–200.

Constantinescu, T. (1986). Schur analysis of positive block-matrices. In *I. Schur Methods in Operator Theory and Signal Processing of Operator Theory Advances and Applications* **18**, 191–206.

Delsarte, P., Genin, Y. and Kamp, Y. (1979). Schur parametrization of positive definite block-toeplitz systems. *SIAM Journal of Applied Math* **36**, 34–46.

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

Dunson, D. B. and Neelon, B. (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics* **59**, 286–295.

Fuller, W. A. (1995). *Introduction to Statistical Time Series*. Wiley-Interscience.

Gallier, J. (2013). Remarks on the Cayley representation of orthogonal matrices and on perturbing the diagonal of a matrix to make it invertible. Available at: *arXiv:math/0606320*.

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.

Giurcanu, M. (2015). A simulation algorithm for non-causal VARMA processes. *Statistics & Probability Letters* **98**, 65–72.

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.

Gourieroux, C. and Jasiak, J. (2016). Filtering, prediction and simulation methods for noncausal processes. *Journal of Time Series Analysis* **37**, 405–430.

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

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.

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

Kaszkurewicz, E. and Bhaya, A. (2000). *Matrix Diagonal Stability in Systems and Computation.* Birkhauser, Boston.

Wise(1956). *Bayesian Multivariate Time-Series Methods in Empirical Macroeconomics.* Now Publishers Inc, Boston.

Koop, G. and Potter, S. M. (2011). Time varying VARs with inequality restrictions. *Journal of Economic Dynamics and Control* **35**, 1126–1138.

Koreisha, S. and Pukkila, T. (1989). Fast linear estimation methods for vector moving average models. *Journal of Time Series Analysis* **10**, 325–339.

Lanne, M. and Saikkonen, P. (2013). Noncausal vector autoregression. *Econometric Theory* **29**, 447–481.

Leonard, T. and Hsu, J. S. J. (2002). Bayesian inference for a covariance matrix. *The Annals of Statistics* **20**, 1669–1696.

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.

Litterman, R. B. (1980). Techniques for forecasting with vector autoregressions. Ph.D. Thesis, University of Minnesota.

Marin, J.-M. and Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics.* Springer, New York.

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.

Mauricio, J. A. (1995). Exact maximum likelihood estimation of stationary vector ARMA models. *Journal of the American Statistical Association* **90**, 282–291.

Mauricio, J. A. (1997). The exact likelihood of a vector autoregressive moving average model. *Applied Statistics* **46**, 157–171.

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.

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.

Nicholls, D. F. and Hall, A. D. (1979). The exact likelihood function of multivariate autoregressive moving average models. *Biometrika* **66**, 259–264.

Nyberg, H., Lanne, M. and Saarinen, E. (2012). Does noncausality help in forecasting economic

time series? *Economics Bulletin* **32**, 2849–2859.

Nyberg, H. and Saikkonen, P. (2014). Forecasting with noncausal VAR model. *Computational Statistics & Data Analysis* **76**, 536–555.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing* **6**, 289–296.

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.

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.

Schneider, H. (1965). Positive operators and an inertial theorem. *Numeriche Mathematik.* **7**, 11–17.

Simionescu, M. (2013). The use of VARMA models in forecasting macroeconomic indicators. *Economics & Sociology* **6**, 94–102.

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

Sims, C. A. (2010). *VAR Tools webpage.* Available at `http://sims.princeton.edu/yftp/VARtools/`.

Stein, P. (1952). Some general theorems on iterants. *Journal of Research of the National Bureau of Standards* **48**, 82-83.

Tunnicliffe-Wilson, G. (1973). The estimation of parameters in multivariate time series models. *Journal of the Royal Statistical Society, B Statistical Methodology* **35**, 76-85.

Tsay, R. (2014). *Multivariate Time Series Analysis With R and Financial Applications*, Wiley.

Velu, R. P., Reinsel, G. C. and Wichern, D. W. (1986). Reduced rank models for multiple time series. *Biometrika* **73**, 105–118.

Whittle, P. (1951). *Hypothesis Testing in Time Series Analysis.* Almquist and Wiksell: Upsala.

Wise, J. (1956). Stationarity conditions for stochastic processes of the autoregressive and moving-average type. *Biometrika* **43**, 215–219.

Zadrozny, P. (1998). An eigenvalue method of undetermined coefficients for solving linear rational expectations models. *Journal of Economic Dynamics and Control* **22**, 1353–1373.

Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore MD 21250, USA.

E-mail: anindya@umbc.edu

Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100, USA.

E-mail: Tucker.S.McElroy@census.gov

Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore MD 21250, USA.

E-mail: pet3@umbc.edu

(Received September 2016; accepted August 2017)