Statistica Sinica

CONSTRAINED ESTIMATION OF CAUSAL INVERTIBLE VARMA

ANINDYA ROY, TUCKER S. MCELROY AND PETER LINTON

University of Maryland Baltimore County and U. S. Census Bureau

Supplementary Material

1. Proofs of Theorems
2. Proposed method in the VAR(1) setting
3. Implementation and additional simulation results

S1. 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 (3.2) such that the coefficients of the polynomial $A = [A_1, \cdots, A_k]$ 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 (3.9). We can see that $\tilde{A}$ satisfies the system of equations

$$\underline{U}_{k-1} = \tilde{A}\underline{U}_{k-1}\tilde{A}' + \tilde{\Sigma} \tag{S1.1}$$

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 $w^* = (w_1^*, \cdots, w_p^*)'$ be any left eigenvector of $\tilde{A}$ corresponding to an eigenvalue λ. 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 (S1.1) 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 $2 \leq t \leq k$,

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

Here terms involving ξ_0, κ_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,

$$C_{t-1} - C_t = \left(U(t) - \xi_{t-1}' \underline{U}_{t-2}^{-1} \, \kappa_{t-1}\right) D_{t-1}^{-1} \left(U(t) - \xi_{t-1}' \underline{U}_{t-2}^{-1} \, \kappa_{t-1}\right)' \geq_L 0. \tag{S1.2}$$

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 $\text{Var}\{X_t', X_{t-1}', \ldots, X_{t-k}'\}$.

For the converse suppose $\underline{U}_{k-1} \in \mathfrak{T}_{++}^{m,k-1}$ exists with $\tilde{S}_k(A, \underline{U}_{k-1}) \in \mathscr{S}_+^{mk}$ and $S_k(A, \underline{U}_{k-1}) \in \mathscr{S}_{++}^m$. 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 $w^* = (w_1^*, \cdots, w_p^*)'$ be any left eigenvector of $\tilde{A}$ corresponding to an eigenvalue λ. Because $S_k(A, \underline{U}_{k-1}) \in \mathscr{S}_{++}^m$, following the arguments given in the proof of Theorem 1, we have $|\lambda| < 1$. $\square$

Proof of Theorem 4. The fact that the map τ 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 ϕ_M and ψ_M are bijections. Let ϕ_M^{-1} be the inverse map of ϕ_M where ϕ^{-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 ψ_M^{-1} denote the inverse map of ψ_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

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

which implies $\phi_M^{-1} \circ \phi_M = id$. Next consider an arbitary element $\underline{U}_k$ of $\mathfrak{T}_{++}^{m,k}(M)$. 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 (3.9) and $\tilde{\Sigma} = \begin{pmatrix} M & 0 \\ 0 & 0 \end{pmatrix}$. The Riccati equations are unique and 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 M, also satisfies the Riccati equations. Therefore $\phi_M \circ \phi_M^{-1} = id$. Also for $\underline{U}_k$ let

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

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 $i = 0, 1, \ldots, (j-1)$. Then at stage j, we have

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

Since V_j, Q_j are obtained under the map ψ_M, from (3.6) and (3.8) we have $V_j^{1/2} Q_j D_{j-1}^{1/2} = U(j)' - \kappa'_{j-1} \underline{U}_{j-2}^{-1} \xi_{j-1}$. Thus $\overleftarrow{\underline{U}}_k = \underline{U}_k$ and hence $\psi_M^{-1} \circ \psi_M = id$. For the converse, let $(V_1, \ldots, V_k, Q_1, \ldots, Q_k) \in (\mathscr{S}_{++}^m)^k \times O(m)^k$ be given. Then $\underline{U}_k = \psi^{-1}(V_1, \ldots, V_k, Q_1, \ldots, Q_k)$ is constructed using Algorithm [VQ]. The inverse images $(\overleftarrow{V}_1, \ldots, \overleftarrow{V}_k, \overleftarrow{Q}_1, \ldots, \overleftarrow{Q}_k)$ are defined using equations (3.6) and (3.8). 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 $j = 1, \ldots, k$. Hence, $\psi_M \circ \psi_M^{-1} = id$. $\square$

S2. Var(1) model

A simpler look at the proposed method using first order VAR

The vector autoregression of order one is the simplest model in terms of parameter estimation and inference in the class of VARMA(p, q) models. To illustrate the pitfalls of the existing likelihood based methods, we analyze the VAR(1) model in the context of reparameterization under causality constraints.

Reparameterization of VAR(1)

Consider an m-dimensional VAR(1) process $\{X_t\}$:

$$X_t = \Phi_1 X_{t-1} + Z_t. \tag{S2.1}$$

For the purpose of writing a likelihood we assume the innovations are normally distributed. The parameters of interest are the $m \times m$ coefficient matrix Φ_1 and the $m \times m$ covariance matrix Σ.

The VAR(1) system is causal iff the roots of $\tilde{\Phi}(z) = z - \Phi_1$ are all within the unit disk $\mathcal{D}$, or equivalently if all the eigenvalues $\lambda_1(\Phi_1), \ldots, \lambda_m(\Phi_1)$ of the matrix Φ_1 are less than one in absolute value. Thus, for the VAR(1), with a slight abuse of notation, we define the parameter space associated with a causal polynomial to be

$$\mathfrak{S}_1^m = \{\Phi_1 \in \mathbb{R}^{m \times m} : |\lambda_j(\Phi_1)| < 1,\ j = 1, \ldots, m\}.$$

The space $\mathfrak{S}_1^m$ forms a submanifold of $\mathbb{R}^{m \times m}$ and is complicated in nature. The space is defined via the eigenvalue restrictions. The eigenvalues are highly nonlinear functions of the elements of Φ_1 and often are not available in explicit form. This description is generally cumbersome to use in any optimization procedure; nor is it easily generalized to cases with $p > 1$.

The key quantity that motivates our reparameterization in the VAR(1) is the observation that $(\Gamma(0), \Sigma, \Phi_1)$ satisfy the discrete algebraic Riccati system,

$$\Gamma(0) = \Phi_1 \Gamma(0) \Phi_1' + \Sigma. \tag{S2.2}$$

It can be shown that any solution of the system for given positive-definite matrices $\Gamma(0) \geq_L \Sigma >_L 0$ will be Schur-Stable. The equation is satisfied by the Yule-Walker solution in the VAR(1) process, $\hat{\Phi}_1 = \Gamma(1)\Gamma(0)^{-1}$, which is known to be Schur-Stable. More generally, the result can be cast in terms of general transformations of symmetric matrices and their relationship to matrix stability analyzed using the Stein transformation, $S(\cdot,\cdot)$, defined in the main text. Thus, for the VAR(1) with stationary variance $\Gamma(0)$ and innovation variance Σ, $S(\Phi_1, \Gamma(0)) = \Sigma$.

Recall that Stein's (1952) result implies that one could characterize $\mathfrak{S}_1^m$ 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 iff $U \geq_L M$ but it need not be a singleton set. In fact, for any $M \in \mathscr{S}_{++}^m$ the entire Schur-Stable class can be generated by the pre-images as U varies over the class of positive definite matrices, i.e., $\mathfrak{S}_1^m = \bigcup_{U \in \mathscr{S}_{++}^m} A_M(U)$. This is immediate since given $M \in \mathscr{S}_{++}^m$ and $A \in \mathfrak{S}_1^m$, one can solve for U as

$$Vec(U) = (I_{m^2} - A \otimes A)^{-1} Vec(M),$$

where $\otimes$ is the kronecker product. Since the pre-images are non-empty iff $U \geq_L M$, we have

$$\mathfrak{S}_1^m = \bigcup_{U \in \mathscr{S}_{++}^m,\, U \geq_L M} A_M(U). \tag{S2.3}$$

For any fixed $M \in \mathscr{S}_{++}^m$, the relation (S2.3) allows us to parameterize the entire Schur-Stable class $\mathfrak{S}_1^m$ in terms of elements of $\mathscr{S}_{++}^m$. However, since the pre-images $A_M(U)$ are not necessarily singletons, we need to introduce additional parameters that can characterize the pre-images uniquely. Before we state our result, we introduce further notation. For $r \leq m$, let $\nu_{r,m}$ denote the Stiefel manifold of $r \times m$ semi-orthogonal matrices. In the special case when r equals m, the set is the orthogonal group $O(m)$ of $m \times m$ orthogonal matrices. Then we have the following result that characterizes the set of Schur-Stable matrices.

Proposition 1. *Let $M \in \mathscr{S}_{++}^m$ be given. Then there exists $A \in \mathfrak{S}_1^m$ and $U \in \mathscr{S}_{++}^m$ such that $S(A, U) = M$ iff there exists $A \in \mathfrak{S}_1^m$ and $U \in \mathscr{S}_{++}^m$ such that $U \geq_L M$ and $A = (U - M)^{1/2} Q U^{-1/2}$ for some $r \times m$ matrix $Q \in \nu_{r,m}$ where $r = rank(U - M)$, $(U - M)^{1/2}$ is a full column rank square root of $(U - M)$ and $U^{-1/2}$ is a square root of U^{-1}.*

Proof of Proposition 1. The "if" part follows immediately from substitution. For the converse note that

$$(U - M)^{1/2}((U - M)^{1/2})' = (AU^{1/2})(AU^{1/2})'.$$

Therefore we can find $Q_1 \in \nu_{r,m}$ such that $(U - M)^{1/2} = AU^{1/2}Q_1'$. $\square$

To see how Proposition 1 provides a characterization of $\mathfrak{S}_1^m$, we fix $M \in \mathscr{S}_{++}^m$ and define $V := V(U) = U - M$ for any $U >_L M$ (for the illustration we only consider the full rank case, i.e. $U >_L M$ but parametrization in the case of $U \geq_L M$ will be immediate from the description.)

Then from Proposition 1, we obtain the alternative parametrization of $A \in \mathfrak{S}_1^m$ in terms of (V, Q) where V is any positive definite matrix and Q is any orthogonal matrix. Note that the number of free parameters in this parameterization is $\binom{m+1}{2}$ for V and $\binom{m}{2}$ for Q. Thus, the total number of free parameters is $\binom{m+1}{2} + \binom{m}{2} = m^2$, the same as that in A. More importantly, the transformation φ taking A to its pre-parameters $(V(A), Q(A))$ is a bijection between $\mathscr{S}_{++}^m \times O(m)$ and $\mathfrak{S}_1^m$.

Consider the map φ and its inverse ϑ (the mappings depend on the choice of M, but we will assume that M is fixed throughout and suppress the dependence for notational simplicity) defined as

$$A \stackrel{\varphi}{\to} (V, Q), \tag{S2.4}$$

$$(V, Q) \stackrel{\vartheta}{\to} A. \tag{S2.5}$$

The formulas for φ and ϑ are

$$\begin{aligned}
V(A) &= \sum_{j\geq 1} A^j M A'^j, \\
Q(A) &= \left(\sum_{j\geq 1} A^j M A'^j\right)^{-1/2} A \left(\sum_{j\geq 0} A^j M A'^j\right)^{1/2}, \\
A(V, Q) &= V^{1/2} Q (V + M)^{-1/2}.
\end{aligned} \tag{S2.6}$$

Proposition 2. *Let φ and ϑ be as defined in (S2.6). Then $\varphi \circ \vartheta = id = \vartheta \circ \varphi$, so that the map is a bijection.*

To establish Proposition 2, we first prove a lemma.

Lemma 1. *If $M \in \mathscr{S}_+$, then $V = \sum_{j\geq 1} A^j M A'^j$ satisfies $V = A(V+M)A'$. Moreover, if $A \in \mathfrak{S}_1^m$, this V is the unique solution to the Riccati equation.*

Proof. The first assertion is trivial algebra. Now suppose there are two solutions, V and $\widetilde{V}$. Then

$$A(V-\widetilde{V})A' = A([V+M]-[\widetilde{V}+M])A' = A[V+M]A'-A[\widetilde{V}+M]A' = V-\widetilde{V},$$

and by taking vec we obtain

$$(I_{m^2} - A \otimes A)\ \mathrm{vec}(V - \widetilde{V}) = 0.$$

Because $A \in \mathfrak{S}_1^m$, the matrix $I_{m^2} - A \otimes A$ is invertible, implying that $\mathrm{vec}(V - \widetilde{V}) = 0$, i.e., $V = \widetilde{V}$. $\square$

Proof of Proposition 2. First consider any A in the domain of φ, which maps to

$$\left(\sum_{j\geq 1} A^j M A'^j, \left(\sum_{j\geq 1} A^j M A'^j \right)^{-1/2} A \left(\sum_{j\geq 0} A^j M A'^j \right)^{1/2} \right).$$

Applying ϑ to this yields

$$\left(\sum_{j\geq 1} A^j M A'^j \right)^{1/2} \left(\sum_{j\geq 1} A^j M A'^j \right)^{-1/2} \Phi_1 \left(\sum_{j\geq 0} A^j M A'^j \right)^{1/2} \left(\sum_{j\geq 0} A^j M A'^j \right)^{-1/2},$$

which equals A; therefore $\vartheta \circ \varphi = \mathrm{id}$. The converse follows from Lemma 1.

Consider any (V, Q) in the domain of ϑ, which is mapped to $V^{1/2}Q$ $(V+M)^{-1/2}$ and we use the abbreviation $\widetilde{A}_1 = V^{1/2}Q(V+M)^{-1/2}$ for convenience. Note that $\widetilde{A}(V+M)\widetilde{A}' = V$ by algebra. Letting $U = V+M$, it then follows that $U = \widetilde{A}U\widetilde{A}_1' + M$. Thus by Stein's result we must have $\widetilde{A}_1 \in \mathfrak{S}_1^m$. Then by Lemma 1 we know that V must be a unique solution to $\widetilde{A}(V+M)\widetilde{A}' = V$. Applying φ we obtain $(\widetilde{V}, \widetilde{Q})$, where

$$\widetilde{V} = \sum_{j\geq 1} \widetilde{A}^j M \widetilde{A}^{j\prime}$$

$$\widetilde{Q} = \widetilde{V}^{-1/2}\widetilde{A}(\widetilde{V}+M)^{1/2}.$$

Noting that $\widetilde{V}$ also solves $\widetilde{A}(\widetilde{V}+M)\widetilde{A}' = \widetilde{V}$ by algebraic verification, Lemma 1 tells us that $\widetilde{V} = V$. Plugging this result back into the formula for $\widetilde{Q}$ yields

$$\widetilde{Q} = V^{-1/2}V^{1/2}Q(V+M)^{-1/2}(V+M)^{1/2} = Q.$$

Therefore $\varphi \circ \vartheta = \mathrm{id}$ as well. $\square$

We end this appendix with an illustration of the advantage of the parameterization in terms of prior specification and by contrasting the approach with common prior specification approaches that disregard the constraints of Schur-stability.

Advantages of constrained prior based on pre-parameterization

Consider a two-dimensional Gaussian VAR(1) process X_t defined by (S2.1) with $m = 2$, and parameters

$$\Phi = \Phi_0 := \begin{pmatrix} \lambda & 0 \\ 2 & \lambda \end{pmatrix} \quad \Sigma = \Sigma_0 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{S2.7}$$

we will take the roots of the process to be near the causal boundary, $\lambda = 1 - n^{-1}$ to illustrate the effect of unrestricted prior specification. Let $(X_1, \ldots, X_{100})$ be a sample of size $n = 100$ from the process. Let $\pi(\Phi, \Sigma)$ be a prior on the parameters. We will assume a priori $\mathrm{Vec}(\Phi) \sim N(\Phi_0, 1_2)$ and independently $\Sigma \sim IW(\Sigma_0, 5 + \kappa)$, where IW denotes the probability density of the inverse-Wishart distribution. The parameterization with $\kappa = 0.5$ produces a reasonably flat prior on Σ with finite prior variance. We center the prior for the coefficient at the true value. The prior belongs to the class of the standard normal-inverse-Wishart (NIW) specification for Bayesian VAR, which includes the popular Minnesota prior (Litterman 1980) as a special case. The Bayes estimator of Φ is then obtained by standard Bayesian computation using the prior and the Gaussian likelihood obtained by conditioning on the initial value. Let $\hat{\lambda}_{1,U}$ denote the largest eigenvalue (in absolute value) of the Bayes estimator of Φ obtained using the NIW type unrestricted prior. For the proposed Bayesian procedure with constrained priors, we induce priors on parameters via priors on the (l, d, s, δ) parameters corresponding to (V, Q, Σ). The exact formulation is given in the general VARMA section, and thus we omit the details here. Let $\hat{\lambda}_{1,C}$ denote the largest eigenvalue (in absolute value) of the Bayes estimator of Φ obtained using the proposed method.

Figure 1 shows the density histogram of the maximum eigenvalues, $\hat{\lambda}_{1,U}$ and $\hat{\lambda}_{1,C}$ based on 400 Monte Carlo replications. The density for the posterior obtained from the unconstrained prior has about 30% posterior mass outside the causal region while the proposed method is concentrated in the causal region. The left tails of distributions for the two estimators are reasonably close, but the unrestricted estimator has posteior eigenvalues with magnitude bigger than one.

S3. Additional Simulation

3.1 Modified Cayley Form

The procedure for recovering skew symmetric S from an orthogonal Q and the corresponding δ value is given below. Derivations of the procedure follow Gallier (2013). Recall $R = [(I_m - S)(I_m + S)^{-1}]^2$ for some skew-symmetric matrix S and $Q = E_\delta R = E_\delta[(I_m - S)(I_m + S)^{-1}]^2$, where $E_\delta = I_m - 2\delta e_1 e_1'$, $\delta \in \{0, 1\}$, and $e_1 = (1, 0, \ldots, 0)'$.

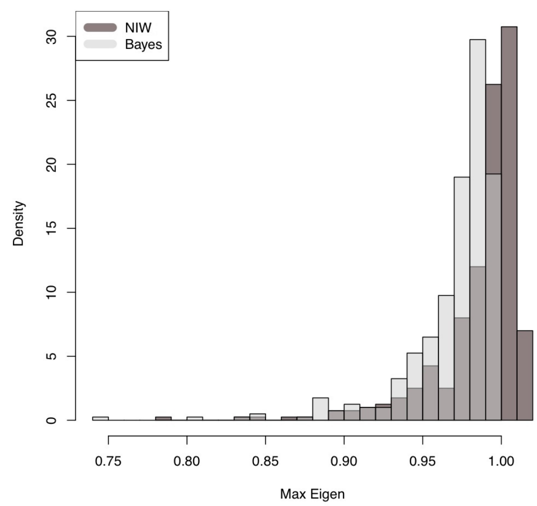

Figure 1: Distribution of the estimated maximum eigen-value (in absolute value) of the coefficient matrix in a VAR(1) model with true values given in (S2.7) and sample size equal to 100. The lighter histogram corresponds to the Bayesian estimator $\hat{\lambda}_{1,C}$ based on the proposed parameterization that constrains the eigenvalue, while the darker histogram is for $\hat{\lambda}_{1,U}$ with standard unconstrained NIW priors on the parameters. The overlapping region between the histograms is shown with an intermediate shade.

1. Recover R from Q as $R = E_\delta Q$. Note that R cannot have an odd number of negative one eigenvalues.

2. Calculate the Schur decomposition of R. $R = PTP'$ where P is orthogonal and T is upper block-triangular with 1×1 or 2×2 blocks along the diagonal. The diagonal 2×2 blocks correspond to complex eigenvalues. The eigenvalues of T are the same as that of R. Because the negative one eigenvalues appear in pairs, we can compute a real square root R or that of T (since $R^{1/2} = PT^{1/2}P'$) as follows:

3. Set $T^{\frac{1}{2}} = I_m$.

4. If diagonal value i and $i-1$ of T are -1, set $\begin{pmatrix} T^{\frac{1}{2}}_{i,i} & T^{\frac{1}{2}}_{i,i+1} \\ T^{\frac{1}{2}}_{i+1,i} & T^{\frac{1}{2}}_{i+1,i+1} \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

5. If diagonal value i and $i-1$ of T are < 1 in absolute value. Let $\theta = arctan(T_{i,i+1}/T_{i,i})$. Set $\begin{pmatrix} T^{\frac{1}{2}}_{i,i} & T^{\frac{1}{2}}_{i,i+1} \\ T^{\frac{1}{2}}_{i+1,i} & T^{\frac{1}{2}}_{i+1,i+1} \end{pmatrix} = \begin{pmatrix} cos(\theta/2) & sin(\theta/2) \\ -sin(\theta/2) & cos(\theta/2) \end{pmatrix}$.

6. Let $R^1 = PT^1P'$.

7. Calculate $S = 2(I_m + R^1)^{-1} - I_m$.

3.2 MLE and Bayesian computation

For MLE we use the BFGS method in R-optim with bounding boxes given in the paper. The initial value is chosen according to the SHIRNK algorithm described below. For Bayesian estimation perform a Metropolis-random walk algorithm with the following recommended priors and jump distribution:

$$
\begin{aligned}
\text{Priors:} & \\
\pi\left(s_1,\ldots,s_{\binom{m}{2}(p+q)}\right) &\propto N(0,1)\\
\pi\left(l_1,\ldots,l_{\binom{m}{2}(p+q+1)}\right) &\propto N(0,1)\\
\pi\left(d_1,\ldots,d_{m(p+q+1)}\right) &\propto N(0,1)\\
\pi\left(z_1,\ldots,z_{p+q}\right) &\propto N(0,1)\\
\text{where } \delta_i &= I[z_i \geq 0] - I[z_i < 0]\\
\text{Jump Distribution:} & \\
\text{For a generic parameter at ith iteration } \eta^{(i)} &\sim N(\eta^{(i-1)}, \psi)
\end{aligned}
$$

Most transformation methods suffer from lack of interpretability in terms of the transformed parameters. For example, a prior on the original parameters invoked via an informative prior on the transformed parameters under a complicated transformation is hardly ever interpretable as a prior on the original parameters. For some intuition about the invoked prior with respect to the original VARMA parameters under the proposed scheme we examine the VAR(1) case. To study the induced distribution on Φ in a VAR(1), set $\eta = (l, d, s)$. Denote the transformation from Φ to η with g, so that $g(a) = f_M(a)$. The recommended prior on η is a multivariate normal $N(\tilde{\eta}, \tilde{\Sigma})$ where $\tilde{\eta} = (0, \ldots, 0)$ and $\tilde{\Sigma} = I$. However, one could specify a general multivariate normal prior, $N(\tilde{\eta}, \tilde{\Sigma})$, on η without adding too much complexity to the MCMC steps but matching it roughly to some informative normal prior on elements of Φ. For example, if the proposed prior on entries of Φ is a multivariate normal with mean vector μ_Φ and covariance Σ_Φ, then using local linearization, the prior on η for the reflection δ corresponding to μ_Φ should be normal with $\tilde{\eta} = g(\mu_\Phi)$ and $\tilde{\Sigma} = \nabla g(\mu_\Phi)'\tilde{\Sigma}_\Phi \nabla g(\mu_\Phi)$. For the other value of δ the proposed independent normal prior on η could be used. Thus, the prior will locally concentrate on μ_Φ.

3.3 Algorithm SHRINK for initial estimation

For implementation of the proposed algorithm we need an initial value that is within the constrained space. For any monic matrix polynomial $A(z)$, define the spectrum of $A(z)$ by

$$\sigma(A) = \{|z| : |A(z)| = |I_m z^k - A_1 z^{k-1} - \cdots - A_k| = 0\} \tag{S3.1}$$

and define $radius(\sigma(A)) = \sup\{|z| : z \in \sigma(A)\}$. Let A be unconstrained estimator, so that possibly $radius\{\sigma(A)\} \geq 1$. Then we convert A to a Schur-stable polynomial using a simple shrinkage technique. Note that

$$\tilde{A}^{(c)}(z) = I_m z^k - cA_1 z^{k-1} - \cdots - c^k A_k \Rightarrow \sigma(\tilde{A}^{(c)}) = c\sigma(A).$$

Then for any $c^{-1} > radius\{\sigma(A)\}$ we have $\tilde{A}^{(c)}(z) \in \mathfrak{S}_k$. For our implementation we use the Hannan-Rissanen estimator (Hannan and Rissanen, 1982) to estimate A, chosen because the procedure is computationally fast and known to be consistent. The shrunk version of the Hannan-Rissanen estimator (autoregressive and/or moving average polynomials are independently shrunk when needed) is used as an initial estimator, which is computationally fast and Schur-stable.

3.4 Additional Simulation results

To evaluate the performance of estimators based on the proposed pre-parameterization we conducted a limited simulation. The models explored are vector autoregression models. In the VARMA setting there are no available causal invertible estimators that can be compared with the proposed estimator. We compared the performance of the proposed estimator with that of the Yule-Walker estimator in VAR(2) in two dimensions ($m = 2$) and VAR(1) in three dimensions ($m = 3$.) We summarize the performance of the estimators using the Monte Carlo root mean squared error based on $N = 500$ Monte Carlo replications of samples of size $n = 100$.

3.5 Three-dimensional VAR(1)

The second setting we consider is a first order three dimensional vector autoregression process. Specifically, we consider the model

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ X_{t,3} \end{pmatrix} = \begin{pmatrix} \Phi_{11} & 0 & 0 \\ 0.1 & 0.5 & 0 \\ \Phi_{31} & 0.4 & 0.8 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \\ X_{t-1,3} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \\ Z_{t,3} \end{pmatrix} \tag{S3.2}$$

where the errors Z_t are independent and identically distributed as $N(0, I_3)$. The different scenarios considered are $\Phi_{31} \in \{0.1, 1\}$ and $\Phi_{11} \in \{-.99, -.95, -0.5, 0.4, 0.9, 0.99\}$. Maximum likelihood estimation is initialized at the Yule-Walker solution. For the MLE optimization box constraints are used on the pre-parameters and the values of the bounds are same as those in the VAR(1) case. We use $N(0, 5)$ priors for the real valued pre-parameters and a *Bernoulli* (0.5) prior for the reflection parameter δ. Since there are 15 parameters in the model, we only report the overall Monte Carlo average of the estimation error for the coefficient matrix, given by $N^{-1} \sum_{j=1}^{N} \|\widehat{\Phi}^{(j)} - \Phi\|$, where $\widehat{\Phi}^{(j)}$ is the estimator of Φ based on the jth Monte Carlo replication. The overall RMSE is reported in Table 1. The results show large efficiency gain for the MLE and the Bayes estimator compared to the Yule-Walker estimator, particularly for processes with roots near the boundary.

3.6 Second order VAR

Next we consider a second order two dimensional vector autoregression process. Specifically, the model is

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} \Phi_{1,11} & 0 \\ 1 & 0.4 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & .45 \end{pmatrix} \begin{pmatrix} X_{t-2,1} \\ X_{t-2,2} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \end{pmatrix} \tag{S3.3}$$

where the errors Z_t are assumed to be i.i.d. $N(0, I_2)$. This particular parameterization provides a one-dimensional parameterization in terms of one of the roots ($\Phi_{1,11}$) of the VAR(2) process and is convenient for illustrating the performance of the estimators as a function of the stability of the process as it changes from very stable to near unit root process. The scenarios considered are $\Phi_{11} \in \{.99, -.95, -.9, -.8, -.6, -.4, -.2, 0, .2, .4, .6, .8, .9, .95, .99\}$. Maximum likelihood estimation is initialized at the Yule-Walker solution. The priors are again chosen in a default manner with $N(0, 5)$ priors for the real valued pre-parameters and $Bernoulli(0.5)$ priors for the reflection parameters. Bayesian computation is done using Metropolis random walk for the real parameters and independent jumps for the reflection parame-

Table 1: Overall RMSE, square-root of $nN^{-1}\sum_{j=1}^{N}\|\hat{\Phi}-\Phi\|^2$, for different estimators of Φ for model (S3.2).

	Φ_{11}	Yule-Walker	Bayes	MLE
$\Phi_{31} = .1$	-0.99	0.207	0.192	0.182
	-0.95	0.204	0.201	0.190
	-0.50	0.241	0.258	0.233
	0.40	0.252	0.259	0.238
	0.90	0.225	0.205	0.199
	0.99	0.227	0.193	0.198
$\Phi_{31} = 1$	-0.99	0.213	0.193	0.183
	-0.95	0.207	0.200	0.190
	-0.50	0.239	0.251	0.229
	0.40	0.241	0.247	0.225
	0.90	0.235	0.197	0.189
	0.99	0.326	0.188	0.191

ters. Due to large number of parameters we only report the Monte Carlo average of the overall estimation error for the autoregressive coefficient matrices $N^{-1}\sum_{j=1}^{N}(\|\hat{\Phi}_1^{(j)}-\Phi_1\|+\|\hat{\Phi}_2^{(j)}-\Phi_2\|)$ where $\widehat{\Phi}_1^{(j)}$ is the estimator of Φ_1 based on the jth Monte Carlo replication and $\widehat{\Phi}_2^{(j)}$ is that for Φ_2. The likelihood based estimators continue to enjoy large efficiency gain over the moment-based estimators in the second order process. The advantage of the proposed parameterization is also seen in terms of the gain in numerical stability of computation near the causal boundary.

Table 2: Overall RMSE, square-root of $nN^{-1}\sum_{j=1}^{N}(\|\hat{\Phi}_1^{(j)} - \Phi_1\| + \|\hat{\Phi}_2^{(j)} - \Phi_2\|)$, for different estimators of Φ_1 and Φ_2 for model (S3.3).

Φ_{11}	Yule-Walker	Bayes	MLE
-0.99	0.602	0.372	0.401
-0.95	0.421	0.368	0.394
-0.90	0.409	0.378	0.400
-0.80	0.387	0.370	0.385
-0.60	0.389	0.366	0.383
-0.40	0.384	0.356	0.374
-0.20	0.380	0.351	0.375
0.00	0.395	0.367	0.391
0.20	0.381	0.362	0.377
0.40	0.403	0.370	0.388
0.60	0.385	0.349	0.363
0.80	0.508	0.354	0.382
0.90	0.597	0.314	0.338
0.95	0.747	0.348	0.355
0.99	0.962	0.359	0.393