SCOTTISH INSTITUTE FOR RESEARCH IN ECONOMICS

SIRE DISCUSSION PAPER

SIRE-DP-2015-06

Large Bayesian VARMAs

Joshua C.C. Chan
Eric Eisenstat
Gary Koop

UNIVERSITY OF STRATHCLYDE

www.sire.ac.uk

Large Bayesian VARMAs*

Joshua C.C. Chan
Australian National University

Eric Eisenstat
University of Bucharest

Gary Koop
University of Strathclyde

September 25, 2014

Abstract: Vector Autoregressive Moving Average (VARMA) models have many theoretical properties which should make them popular among empirical macroeconomists. However, they are rarely used in practice due to over-parameterization concerns, difficulties in ensuring identification and computational challenges. With the growing interest in multivariate time series models of high dimension, these problems with VARMAs become even more acute, accounting for the dominance of VARs in this field. In this paper, we develop a Bayesian approach for inference in VARMAs which surmounts these problems. It jointly ensures identification and parsimony in the context of an efficient Markov chain Monte Carlo (MCMC) algorithm. We use this approach in a macroeconomic application involving up to twelve dependent variables. We find our algorithm to work successfully and provide insights beyond those provided by VARs.

Keywords: VARMA identification, Markov Chain Monte Carlo, Bayesian, stochastic search variable selection

JEL Classification: C11, C32, E37

*Gary Koop is a Senior Fellow at the Rimini Center for Economic Analysis. Emails: joshuacc.chan@gmail.com, eric.eisenstat@gmail.com and gary.koop@strath.ac.uk

1 Introduction

Vector autoregressions (VARs) have been extremely popular in empirical macroeconomics and other fields for several decades (e.g. beginning with early work such as Sims, 1980, Doan, Litterman and Sims, 1984 and Litterman, 1986 with recent examples being Korobilis, 2013 and Koop, 2014). Until recently, most of these VARs have involved only a few (e.g. two to seven) dependent variables. However, VARs involving tens or even hundreds of variables are increasingly popular (see, e.g., Banbura, Giannone and Reichlin, 2010, Carriero, Clark and Marcellino, 2011, Carriero, Kapetanios and Marcellino, 2009, Giannone, Lenza, Momferatou and Onorante, 2010 and Koop, 2013, and Gefang, 2014). Vector autoregressive moving average models (VARMAs) have enjoyed less popularity with empirical researchers despite the fact that theoretical macroeconomic models such as dynamic stochastic general equilibrium models (DSGEs) lead to VMA representations which may not be well approximated by VARs, especially parsimonious VARs with short lag lengths. Papers such as Cooley and Dwyer (1998) point out the limitations of the structural VAR (SVAR) framework and suggest VARMA models as often being more appropriate. For instance, Cooley and Dwyer (1998) conclude "While VARMA models involve additional estimation and identification issues, these complications do not justify systematically ignoring these moving average components, as in the SVAR approach." There is, thus, a strong justification for the empirical macreconomist's toolkit to include VARMAs.

VARs are commonly used for forecasting. But, for the forecaster, too, there are strong reasons to be interested in VARMAs. The univariate literature contains numerous examples in finance and macroeconomics where adding MA components to AR models improves forecasting (e.g. Chan, 2013). But even with multivariate macroeconomic forecasting some papers (e.g. Athanasopoulos and Vahid, 2008) have found that VARMAs forecast better than VARs. Theoretical econometric papers such as Lutkepohl and Poskitt (1996) also point out further advantages of VARMAs over VARs.

Despite these advantages of VARMA models, they are rarely used in practice. There are three main reasons for this. First, there are difficult identification problems to be overcome. Second, VARMAs are parameter rich models which can be over-parameterized (an especially important concern in light of the growing interest in large dimensional models as is evinced in the growing large VAR literature). And, largely due to the first two problems, they can be difficult to estimate. This paper develops methods for estimating VARMAs which address all these concerns.

The paper is organized in the following sections. Section 2 briefly describes the econometric theory of VARMAs paying particular attention to different parameterizations of the VARMA including the expanded form (which is used in the main part of our MCMC algorithm) and the canonical echelon form (which is used in our treatment of identification). Section 3 describes our approach which uses Bayesian methods and a hierarchical prior to jointly select identification restrictions and ensure shrinkage in the resulting model. An MCMC algorithm which implements our approach is developed. Section 4 investigates how well our approach works in practice through an artificial data exercise and a substantive macroeconomic application using VARMAs containing up to 12 variables. We find that our methods are computationally feasible and lead to inference on

parameters and impulse responses that are more reasonable and estimated more accurately than alternative approaches, especially in the larger VARMAs of interest in modern macroeconomics.

2 The Econometrics of VARMAs

2.1 The Semi-structural VARMA

Consider the n dimensional multivariate time series $\mathbf{y}_t, t = -\infty, \ldots, \infty$ and begin with the semi-structural form of the VARMA(p, q):

$$\mathbf{B}_0\mathbf{y}_t = \sum_{j=1}^{p} \mathbf{B}_j\mathbf{y}_{t-j} + \sum_{j=1}^{q} \mathbf{\Theta}_j\boldsymbol{\epsilon}_{t-j} + \mathbf{\Theta}_0\boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{1}$$

or, in terms of matrix polynomial lag operators,

$$\mathbf{B}(L)\mathbf{y}_t = \mathbf{\Theta}(L)\boldsymbol{\epsilon}_t,$$

and assume stationarity and invertibility. For future reference, denote the elements of the VAR and VMA parts of the model as $\mathbf{B}(L) = [\beta_{ki}(L)]$ and $\mathbf{\Theta}(L) = [\theta_{ki}(L)]$ for $i, k = 1, .., n$.

The theoretical motivation for the VARMA arises from the Wold decomposition:

$$\mathbf{y}_t = \mathbf{K}(L)\boldsymbol{\epsilon}_t, \tag{2}$$

where $\mathbf{K}(L)$ is generally an infinite degree polynomial operator. Specifically, it can be shown that any such rational transfer function $\mathbf{K}(L)$ corresponds to the existence of two finite degree operators $\mathbf{B}(L)$ and $\mathbf{\Theta}(L)$ such that

$$\mathbf{B}(L)\mathbf{K}(L) = \mathbf{\Theta}(L).$$

Thus, the VARMA(p, q) is an exact finite-order representation of any multivariate system that can be characterized by a rational transfer function. When $\mathbf{K}(L)$ is not rational, the VARMA(p, q) can provide an arbitrarily close approximation. Moreover, an important advantage of the VARMA class is that, unlike VARs or pure VMAs, it is closed under a variety of transformations on $\mathbf{y}_t$, including linear operations and subsets.

The practical problem in having both AR terms with MA terms, however, is that an alternative VARMA with coefficients $\mathbf{B}^\dagger(L) = \mathbf{C}(L)\mathbf{B}(L)$ and $\mathbf{\Theta}^\dagger(L) = \mathbf{C}(L)\mathbf{\Theta}(L)$ will lead to the same Wold representation. The VARMA(p, q) representation, therefore, is in general not unique. However, there are two reasons why a unique representation is desirable in practice: parsimony and identification. The first reason concerns both frequentist and Bayesian approaches. If $\mathbf{B}(L)$ and $\mathbf{\Theta}(L)$ contain redundancies, then the resulting model may lead to poor forecast performance and imprecise impulse response functions. For researchers working with larger VARMAs such over-parameterization concerns can become severe. For instance, in our empirical work, we use as an estimating model the 12-variate VARMA with four lags (and an intercept in each equation). Even imposing

$\mathbf{B}_0 = \mathbf{\Theta}_0 = \mathrm{I}$ leaves 1,242 parameters (including error covariances) to estimate. With macroeconomic data sets containing a few hundred observations, it will be very hard to obtain precise inference for all these parameters in the absence of an econometric method which ensures parsimony or shrinkage.

The second reason (lack of identification) may be less important for the Bayesian who is only interested in forecasting or in identified functions of the parameters such as impulse responses. That is, given a proper prior a well-defined posterior will exist even in a non-identified VARMA. However, the role of the prior becomes important in such cases and carelessly constructed priors can lead to deficient inference for the Bayesian. For frequentists, however, a lack of identification is a more substantive problem, precluding estimation.

How does one obtain a unique VARMA representation? There are generally two major steps:

The first step is to eliminate common roots in $\mathbf{B}(L), \mathbf{\Theta}(L)$ such that only $\mathbf{C}(L)$ with a constant determinant is possible. In this case, the operators $\mathbf{B}(L), \mathbf{\Theta}(L)$ are said to be left coprime and $\mathbf{C}(L)$ unimodular. For the univariate case, it is sufficient to achieve uniqueness and corresponds in practical terms to specifying minimal orders p, q. For a multivariate process, however, this is not enough and a second step is required. That is, even if we impose $\mathbf{B}_0 = \mathbf{\Theta}_0 = \mathrm{I}$, there may still exist $\mathbf{C}(L) \neq \mathrm{I}$ that preserves this restriction for an alternative set of left coprime operators $\mathbf{B}^\dagger(L), \mathbf{\Theta}^\dagger(L)$. A common example is

$$\mathbf{C}(L) = \begin{pmatrix} 1 & c(L) \\ 0 & 1 \end{pmatrix}.$$

Clearly, $\det \mathbf{C}(L) = 1$ and for any $\mathbf{B}(L), \mathbf{\Theta}(L)$, the transformations $\mathbf{B}^\dagger(L) = \mathbf{C}(L)\mathbf{B}(L)$ and $\mathbf{\Theta}^\dagger(L) = \mathbf{C}(L)\mathbf{\Theta}(L)$ lead to $\mathbf{B}_0^\dagger = \mathbf{\Theta}_0^\dagger = \mathrm{I}$.

This implies that the elements of $\mathbf{B}(L), \mathbf{\Theta}(L)$ are not identified for estimation purposes. One approach to achieving identification relies on the assumption that the matrix $[\mathbf{B}_p : \mathbf{\Theta}_q]$ has full row rank, and indeed, when this holds then $\mathbf{B}_0 = \mathbf{\Theta}_0 = \mathrm{I}$ induces a unique representation (e.g., Hannan, 1976). In practice, one could try to explicitly enforce $[\mathbf{B}_p : \mathbf{\Theta}_q]$ to have full row rank, but that may not be desirable in many applications. The full row rank condition will likely not be satisfied by most data generating processes (DGPs) in practice (Lütkepohl and Poskitt, 1996). Therefore, forcing it in an estimation routine would likely result in mis-specification and an alternative second step would be required to achieve uniqueness when $[\mathbf{B}_p : \mathbf{\Theta}_q]$ is rank deficient.

The more general approach that we follow involves imposing exclusion restrictions on elements of $\mathbf{B}(L), \mathbf{\Theta}(L)$ such that only $\mathbf{C}(L) = \mathrm{I}$ is possible. It turns out that when such zero restrictions are applied according to a specific set of rules, it is possible to achieve a unique VARMA representation corresponding to a particular rational $\mathbf{K}(L)$. This leads to the echelon form which we will use as a basis for our approach to identification.

2.2 The Echelon Form for the VARMA

The echelon form involves a particular set of restrictions on the semi-structural VARMA. The derivation of the echelon form is based on Kronecker index theory which shows that

every $\mathbf{K}(L)$ in (2) is associated with a unique set of indices $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_n)$, which can be directly related to the VARMA operators $\mathbf{B}(L), \boldsymbol{\Theta}(L)$. Identification is achieved by imposing restrictions on the VARMA coefficients in (1) according to so-called Kronecker indices $\kappa_1, \ldots, \kappa_n$, with $0 \leq \kappa_i \leq p^*$, where $p^* = \max\{p, q\}$.

To explain further the identifying restrictions in the echelon form note that, without loss of generality, we can denote the VARMA(p, q) as VARMA(p^*, p^*). Then any VARMA(p^*, p^*) can be represented in echelon form by setting $\mathbf{B}_0 = \boldsymbol{\Theta}_0$ to be lower triangular with ones on the diagonal and applying the exclusion restrictions defined by $\boldsymbol{\kappa}$ to $\mathbf{B}_0, \ldots, \mathbf{B}_{p^*}, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{p^*}$. The latter restrictions impose on $[\mathbf{B}(L) : \boldsymbol{\Theta}(L)]$ a maximal degree of each row i equivalent to κ_i. A VARMA in echelon form is denoted VARMA$_E(\boldsymbol{\kappa})$ and details regarding the foregoing restrictions are discussed in many places. The key theoretical advantage of the echelon form is that, given $\boldsymbol{\kappa}$, it provides a way of constructing a parsimonious VARMA representation for $\mathbf{y}_t$. A by-product of this is that the unrestricted parameters are identified. At the same time, every conceivable VARMA can be represented in echelon form. The formal definition of the echelon form is given, e.g., in Lutkepohl, 2005, page 453 as:

Definition:

The VARMA representation in (1) is in echelon form if the VAR and VMA operators are left coprime and satisfy the following conditions.

The VAR operator is restricted as (for $k, i = 1, \ldots, n$):

$$\begin{array}{l} \beta_{kk}(L) = 1 - \sum_{j=1}^{p_k} \beta_{kk,j} L^j \text{for } k = 1, \ldots, n \\ \beta_{ki}(L) = -\sum_{j=p_k - p_{ki}+1}^{p_k} \beta_{ik,j} L^j \text{ for } k \neq i \end{array},$$

where

$$p_{ki} = \begin{cases} \min(p_k + 1, p_i) \text{ for } k \geq i \\ \min(p_k, p_i) \text{ for } k < i \end{cases}.$$

The VMA operator is restricted as (for $k, i = 1, \ldots, n$):

$$\theta_{ki}(L) = \sum_{j=0}^{p_k} \theta_{ki,j} L^j \text{ and } \boldsymbol{\Theta}_0 = \mathbf{B}_0.$$

The row degrees of each polynomial are $p_1, \ldots, p_n$. In the echelon form the row degrees are the Kronecker indices which we label $\kappa_1, \ldots, \kappa_n$.

We specify a distinction between row degrees $(p_1, \ldots, p_n)$ and Kronecker indices $(\kappa_1, \ldots, \kappa_n)$ since this plays a role in our MCMC algorithm. In this, at one stage we work with a VARMA that simply has row degrees $p_1, \ldots, p_n$, but is otherwise unrestricted. That is, it does not impose the additional restrictions (defined through p_{ki}) required to put the VARMA in echelon form.

As an example of the echelon form, consider a bivariate VARMA$(1, 1)$, denoted as

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} \begin{pmatrix} \epsilon_{1,t-1} \\ \epsilon_{2,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix}. \tag{3}$$

If it is known that $\beta_{21} = \beta_{22} = \theta_{21} = \theta_{22} = 0$, then $y_{2,t} = \epsilon_{2,t}$ and β_{12} is not separately identified from θ_{12}. To achieve identification in this case, it is sufficient to restrict either

$\beta_{12} = 0$ or $\theta_{12} = 0$. However, knowing that $y_{2,t} = \epsilon_{2,t}$ implies that the Kronecker indices of the system are $\kappa_1 = 1, \kappa_2 = 0$. Converting (3) to a VARMA$_E(1,0)$ yields

$$\begin{pmatrix} 1 & 0 \\ \beta_0 & 1 \end{pmatrix} \begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} \beta_{11} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \theta_{11} & \theta_{12} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \epsilon_{1,t-1} \\ \epsilon_{2,t-1} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \beta_0 & 1 \end{pmatrix} \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix}.$$

Therefore, the rules associated with the echelon form automatically impose the identifying restriction $\beta_{12} = 0$.

The key challenge of applying the echelon form methodology in practice is specifying $\boldsymbol{\kappa}$. After all, any unrestricted VARMA is also a VARMA$_E$ for a particular set of Kronecker indices. The problem is that whenever a particular κ_i is over-specified, the resulting VARMA$_E$ is unidentified; whenever it is under-specified, the VARMA$_E$ is misspecified. Therefore, to exploit the theoretical advantages that the VARMA$_E$ provides, the practitioner must choose the Kronecker indices correctly.

The standard frequentist approach to specifying and estimating VARMA models, in consequence, can be described as consisting of three steps:

1. estimate the Kronecker indices, $\hat{\boldsymbol{\kappa}}$;
2. estimate model parameters of the VARMA$_E(\hat{\boldsymbol{\kappa}})$;
3. reduce the model (e.g. using hypothesis testing procedures to eliminate insignificant parameters).

It is important to emphasize that the order of the above steps is crucial. Specifically, step 2 cannot be reasonably performed without completing step 1 first. To appreciate the difficulties with implementing step 1, however, consider performing a full search procedure over all possible Kronecker indices for an n-dimensional system. This would require setting a maximum order $\kappa_{\max}$, estimating $(\kappa_{\max} + 1)^n$ echelon form models implied by each combination of Kronecker indices and then applying some model selection criterion to select the optimal $\boldsymbol{\kappa}$. Given the difficulties associated with maximizing a VARMA$_E$ likelihood, even conditional on a perfectly specified $\boldsymbol{\kappa}$, one cannot hope to complete such a search in a reasonable amount of time (i.e. even a small system with $n = 3$ and $\kappa_{\max} = 5$ would require 1024 Full Information Maximum Likelihood (FIML) routines). Moreover, many of the combinations of $\kappa_1, \ldots, \kappa_n$ that a full search algorithm would need to traverse inevitably result in unidentified specifications, thus plaguing the procedure with exactly the problem that it is designed to resolve.

To handle this difficulty, abbreviated search algorithms relying on approximations are typically employed. Poskitt (1992) provides one particularly popular approach. First, it takes advantage of some special features that arise if the Kronecker indices are re-ordered from smallest to largest such that the number of model evaluations is greatly reduced. Second, it involves a much simpler estimation routine for each evaluation step—i.e., a closed form procedure for consistently (though less efficiently than FIML) estimating the free parameters of a VARMA$_E(\boldsymbol{\kappa})$. These two features also alleviate (although do not eliminate) the problem of needing to estimate unidentified specifications over the course of the search. As a result, consistent estimates of the Kronecker indices are obtained.

However, the implementation also relies on a number of approximations. First, like all existing Kronecker search algorithms, Poskitt (1992) begins by estimating residuals from a long VAR. These are then treated as observations in subsequent least squares estimation routines, which are used to compute information criteria for models of alternative Kronecker structures. Based on the model comparison, the search algorithm terminates when a local optimum is reached. In small samples, therefore, the efficiency of this approach will depend on a number of manual settings and may often lead to convergence difficulties in the likelihood maximization routines implemented at the second stage (for further discussion, see Lutkepohl and Poskitt, 1996).

Consequently, the procedure does not really overcome the basic hurdle: if the $\hat{\boldsymbol{\kappa}}$ obtained in small samples incorrectly describes the underlying structure of the Kronecker indices (as reliable as it may be asymptotically), the VARMA$_E(\hat{\boldsymbol{\kappa}})$ specified in step 2 may ultimately be of little use in resolving the specification and identification issues associated with the unrestricted VARMA.

Recently, Dias and Kapetanios (2013) have developed a computationally-simpler iterated ordinary least squares (OLS) estimation procedure for estimating VARMAs. They prove its consistency and, although it is less efficient than the maximum likelihood estimator (MLE), it has the advantage that it works in places where the MLE does not. In fact, the authors conclude (page 22) that "the constrained MLE algorithm is not a feasible alternative for medium and large datasets due to its computational demand." For instance, they report that their Monte Carlo study which involved 200 artificial generated data sets of 200 observations each from an 8 dimensional VARMA took almost one month of computer time. Their iterated OLS procedure is an approximate method, but the authors show its potential to work with larger VARMAs such as those considered in the present paper. However, their method does run into the problem that it can often fail to converge when either the sample size is small or the dimension of the VARMA is large. For instance, their various Monte Carlo exercises report failure to convergence rates from 79% to 97% for VARMAs with 10 dependent variables and T=150. These results are generated with VARMA(1,1) models and would, no doubt, worsen with longer lag lengths such as those considered in the present paper. These high failure to converge rates are likely due to the fact that, with many parameters to estimate and relatively little data to inform such estimates, likelihood functions (or approximations to them) can be quite flat and their optima difficult to find. This motivates one theme of our paper: use of carefully selected shrinkage through a Bayesian prior is useful in producing sensible (and computationally feasible) results in large VARMA models.

It is not difficult to see why applied macroeconomists have rarely used these frequentist procedures for estimating VARMAs. To extend them to models with stochastic volatility or a similar feature commonly incorporated in modern VARs seems extremely difficult. Shortly, we will develop a Bayesian method which jointly goes through the three steps listed above in the context of a single MCMC algorithm and allows for many extensions (e.g. adding stochastic volatility) in a straightforward fashion. Before we do this, we describe an alternative way of parameterizing the VARMA which is used in our MCMC algorithm.

2.3 The Expanded Form for the VARMA

Papers such as Metaxoglou and Smith (2007) and Chan and Eisenstat (2014) adopt an alternative way of parameterizing the VARMA called the expanded VARMA form which proves useful for computational purposes. The expanded VARMA form, which provides an equivalent representation, can be written as:

$$\mathbf{B}_0\mathbf{y}_t = \sum_{j=1}^{p} \mathbf{B}_j\mathbf{y}_{t-j} + \sum_{j=0}^{p} \mathbf{\Phi}_j\mathbf{f}_{t-j} + \boldsymbol{\eta}_t, \tag{4}$$

where $\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega})$ and $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$ are independent, $\mathbf{\Phi}_0$ is a lower triangular matrix with ones on the diagonal, $\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p$ are coefficient matrices, and $\mathbf{\Omega}, \mathbf{\Lambda}$ are diagonal. Since the parameters in the echelon form VARMA or the semi-structural VARMA can be recovered from the expanded VARMA parameters, it is clear that estimating the latter is sufficient to estimate the former. Our MCMC algorithm draws from the expanded VARMA form and then transforms draws to the echelon form.

Chan and Eisenstat (2014) provide an extensive discussion of the expanded VARMA form and its use in building MCMC algorithms. To summarize briefly, the expanded form introduces n additional parameters, which are not fully identified even with the echelon form restrictions imposed. However, this expansion of the parameter space does not require restrictive priors, nor does it impair sampling efficiency. In fact, because it leads to a straightforward linear state-space model, one can readily take advantage of a number of existing computational methods to construct fast sampling algorithms. Moreover, working directly with the expanded form and transforming the draws *ex post* to recover the original VARMA parameters circumvents the need to impose invertibility restrictions in the course of the MCMC. Instead, this is easily implemented in the ex post processing of draws, i.e. in transforming $\mathbf{\Phi}_0, \ldots, \mathbf{\Phi}_q, \mathbf{\Omega}, \mathbf{\Lambda}$ to $\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_q, \mathbf{\Sigma}$. A straightforward algorithm for implementing the latter is detailed in Section 2 of Chan and Eisenstat (2014).

3 Bayesian Inference in VARMA Models

3.1 The Existing Bayesian Literature

Previously, we have drawn a distinction between the related concepts of parsimony and identification. Identification can be achieved by selecting the correct Kronecker indices (which imply certain restrictions on a semi-structural VARMA model). Parsimony is a more general concept, involving either setting coefficients to zero (or any constant) or shrinking them towards zero. So identification can be achieved through parsimony (i.e. selecting the precise restrictions implied by the Kronecker indices in the context of an unidentified VARMA model), but parsimony can involve imposing other restrictions on a non-identified model or imposing restrictions beyond that required for identifying the model.

In this sense, the Bayesian literature breaks into two groups. The first consists of papers which estimate VARMA models, possibly taking into account parsimony considerations. Good examples of this literature are Ravishanker and Ray (1997) and Chan and

Eisenstat (2014). The second consists of papers which explicitly address identification issues. The key references in this strand of the literature is Li and Tsay (1998). Since a key focus of our paper lies in identification, we will focus on this paper.

Li and Tsay (1998) specify a model similar to (1), i.e., with $\mathbf{\Theta}_0$ lower triangular but not equal to $\mathbf{B}_0$ and a diagonal $\mathbf{\Sigma}$ and work with the echelon form, attempting to jointly estimate the VARMA parameters with the Kronecker indices (as we do in the present paper). This is done through the use of a hierarchical prior for the coefficients which is often called a stochastic search variable selection (SSVS) prior (although other terminologies exist). Before describing Li and Tsay's algorithm, we briefly introduce the idea underlying SSVS in a generic context. Let α be a parameter. SSVS specifies a hierarchical prior (i.e. a prior expressed in terms of parameters which in turn have a prior of their own) which is a mixture of two Normal distributions:

$$\alpha \,|\, \gamma \sim (1-\gamma)\mathcal{N}(0, \tau_0^2) + \gamma\mathcal{N}(0, \tau_1^2), \tag{5}$$

where γ is a dummy variable. Thus, if $\gamma = 1$ then the prior for α is given by the second Normal and if $\gamma = 0$ it is given by the first Normal. The prior is hierarchical since γ is treated as an unknown parameter and estimated in a data-based fashion. The aspect which allows for prior shrinkage and variable selection arises by choosing the first prior variance, τ_0^2, to be "small"(so that the coefficient is shrunk so as to be close to zero) and the second prior variance, τ_1^2, to be "large"(implying a relatively noninformative prior for the corresponding coefficient). An SSVS prior of this sort, which we shall call "soft SSVS", has been used by many researchers. For instance, George, Sun and Ni (2008) and Koop (2013) use it with VARs and Li and Tsay (1998) adopt something similar. An extreme case of the SSVS prior arises if the first Normal in (5) is replaced by a point mass at zero. This we will call "hard SSVS". It was introduced in Kuo and Mallick (1997) and used with VARs by Korobilis (2013) and others.

Li and Tsay (1998) specify soft SSVS priors on the VAR and VMA coefficients of a VARMA. The ingenuity of this approach is that it combines in practical terms the two related concepts of identification and parsimony. The authors enforce the echelon form through this framework by imposing certain deterministic relationships between the SSVS indicators (see section 4 of Li and Tsay, 1998, for more details). Based on this, they devise an MCMC algorithm that cycles through n individual (univariate) ARMAX equations. The ith ARMAX equation is obtained by treating the observations $\{y_{j,t}\}$ for $j = 1, \ldots, i-1$, $t = 1, \ldots, T$ and the computed errors $\{\epsilon_{j,t}\}$ for $j \neq i, t = 1, \ldots, T$ as exogenous regressors. SVSS indicators are then updated conditional on draws of the coefficients and subject to the deterministic relationships implied by the echelon form. In consequence, draws of the Kronecker indices (which can be recovered from draws of the SSVS indicators) are simultaneously generated along with the model parameters.

Their algorithm, however, entails a significant degree of complexity both in terms of programming and computation. A pass through each equation requires reconstructing VARMA errors (i.e. based on previous draws of parameters pertaining to other equations) and sampling three parameter blocks: (i) the autoregressive and "exogenous" variable coefficients, (ii) the error variance, and (iii) the moving average parameters. The latter entails a non-trivial Metropolis-Hastings step, and all must be repeated n times for every

sweep of the MCMC routine. Evidently, the complexity of this algorithm grows rather quickly with the size of the system, and in their applications, only systems with $n = 3$ and $\kappa_{\max} \leq 3$ are considered. The run times reported for even these small systems are measured in hours.

Relative to Li and Tsay (1998) our algorithm shares the advantage of jointly estimating Kronecker indices and model parameters, thus ensuring parsimony and identification. However, we argue that ours is a more natural specification, which also provides great computational benefits and allows us to work with the large Bayesian VARMAs of interest to empirical macroeconomists. First, by using the expanded form discussed in subsection 1.4, we are able to work with a familiar, linear state-space model. Conditional on the Kronecker indices, computation is fast and efficient even for large n. Moreover, this representation enables us to analytically integrate out the coefficients $\{\mathbf{B}_j\}$ and $\{\mathbf{\Phi}_j\}$ when sampling the Kronecker indices. The efficiency gains from this are particularly important as n increases because the size of each $\mathbf{B}_j$ and $\mathbf{\Phi}_j$ grows quadratically with n. In fact, this added efficiency together with the reduced computational burden is precisely what allows us to estimate an exact echelon form VARMA for large systems. The details are provided in the following subsection.

Chan and Eisenstat (2014) develop an MCMC algorithm on the expanded form of the VARMA. However, it does not deal with identification using the echelon form as is done in the present paper. Nor does it deal with the challenges involved with large VARMAs (e.g. it does not develop shrinkage priors such as those we introduce shortly). Nevertheless, the MCMC algorithm developed in Chan and Eisenstat (2014) is the building block that we extend in the present paper when we derive an MCMC algorithm for the canonical echelon form.

3.2 Our Approach to Bayesian Inference in VARMAs

Our approach to Bayesian inference is based on the ideas that identification is achieved in the echelon form (i.e. through estimating $\boldsymbol{\kappa}$ in the VARMA$_E(\boldsymbol{\kappa})$), but computation is more easily done in the VARMA(p, q) model in the expanded form (see also Chan and Eisenstat, 2014). Thus, our MCMC algorithm works in the latter, but draws are transformed to the echelon form. We also treat $\boldsymbol{\kappa}$ as a vector of unknown parameters and draw it in our algorithm. Parsimony and identification are achieved using SSVS priors.

In particular, we parameterize the echelon form by row degrees $p_1, \ldots, p_n$ and impose two sets of restrictions: those implied by the row degrees and those resulting from the additional shrinking of model parameters. As will be made clear in this subsection, the echelon form is enforced by imposing a certain relationship between the two types of restrictions. All these elements are introduced in the model through a unified hierarchical SSVS prior.

Since row degree restrictions are especially important for identification, we always use hard SSVS for these (i.e. the restrictions implied by a choice for $p_1, \ldots, p_n$ and, thus, $\boldsymbol{\kappa}$ are imposed exactly). Restrictions on the remaining parameters are partly used for identification and partly to achieve additional parsimony (i.e., by further restricting parameters which remain in the VARMA$_E(\boldsymbol{\kappa})$). For these, the researcher may wish to use either soft or hard SSVS and, in this paper, we allow for both. With some abuse of

terminology, we will call the prior which uses hard SSVS to achieve identification and soft SSVS to achieve parsimony the "soft SSVS prior" and the prior which imposes hard SSVS throughout the "hard SSVS prior". In what follows, we describe the main features of our approach, paying particular attention to the SSVS priors on the VARMA coefficients. Complete details on the priors for the remaining parameters are given in Appendix A. Complete details of our MCMC algorithm are given in Appendix B.

Consider the expanded form VARMA given in (4) for which the VARMA coefficients are parameterized in terms of $\mathbf{B}_i$ and $\mathbf{\Phi}_i$. Let the individual coefficients in these matrices be denoted $B_{i,jk}$ and $\Phi_{i,jk}$, respectively. Here we describe the soft SSVS implementation with $\tau^2_{0,ijk} \ll \tau^2_{1,ijk}$ (the hard SSVS implementation will be the same except there is no $\tau^2_{0,ijk}$, but instead a point mass at zero is used) which is given by

$$\begin{aligned}
\left(B_{i,jk} \,|\, \gamma^{B,R}_{ijk}, \gamma^{B,S}_{ijk}\right) \sim{}& \left(1-\gamma^{B,R}_{ijk}\right) \mathbb{1}(B_{i,jk}=0) \\
&+\gamma^{B,R}_{ijk}\left(\left(1-\gamma^{B,S}_{ijk}\right)\mathcal{N}(0,\tau^2_{0,ijk}) + \gamma^{B,S}_{ijk}\mathcal{N}(0,\tau^2_{1,ijk})\right), \\
\left(\Phi_{i,jk} \,|\, \gamma^{\Phi,R}_{ijk}, \gamma^{\Phi,S}_{ijk}\right) \sim{}& \left(1-\gamma^{\Phi,R}_{ijk}\right) \mathbb{1}(\Phi_{i,jk}=0) \\
&+\gamma^{\Phi,R}_{ijk}\left(\left(1-\gamma^{\Phi,S}_{ijk}\right)\mathcal{N}(0,\tau^2_{0,ijk}) + \gamma^{\Phi,S}_{ijk}\mathcal{N}(0,\tau^2_{1,ijk})\right),
\end{aligned} \tag{6}$$

where $\mathbb{1}(\cdot)$ is the indicator function. In this setup, $\gamma^{B,R}_{ijk}, \gamma^{\Phi,R}_{ijk} \in \{0,1\}$ are indicators determined completely by the row degrees: $\gamma^{B,R}_{ijk} = \gamma^{\Phi,R}_{ijk} = 1$ *iff* $0 < j \leq \rho_i$ or $j = 0, i < k$. Furthermore, $\gamma^{B,S}_{ijk}, \gamma^{\Phi,S}_{ijk} \in \{0,1\}$ are the indicators related to the SSVS mechanism for the remaining coefficients not restricted by the row degrees.

Using the definition of the echelon form in subsection 1.3, define a mapping $\mathcal{E}(\boldsymbol{\kappa})$ from the Kronecker indices to a set of indicators on the coefficients. We can use this mapping in the construction of a set of restriction indicators for the echelon form: $\boldsymbol{\gamma}^E = \{\gamma^{B,E}_{ijk}, \gamma^{\Phi,E}_{ijk}\} = \mathcal{E}(p_1,\ldots,p_n)$. The echelon form can be imposed by specifying the prior on $\gamma^{B,S}_{ijk}$ conditional on $p_1,\ldots,p_n$ as

$$\Pr\left(\gamma^{B,S}_{ijk} = 1 \,|\, p_1,\ldots,p_n\right) = \begin{cases} 0 & \text{if } \gamma^{B,E}_{ijk} = 0, \gamma^{B,R}_{ijk} \neq 0 \\ 0.5 & \text{otherwise} \end{cases}. \tag{7}$$

For the indicators on the elements of $\mathbf{\Phi}_j$ we set $\Pr\left(\gamma^{\Phi,S}_{ijk} = 1\right) = 0.5$.[1] We further set uniform priors on $p_1,\ldots,p_n$, which induce a prior on $\boldsymbol{\gamma}^R$, and by implication, a uniform prior on the Kronecker indices. Our MCMC algorithms provide draws of $p_1,\ldots,p_n$, and under the prior specification (7), these are equivalent to draws of the Kronecker indices $\kappa_1,\ldots,\kappa_n$. Parameters of interest in terms of (1) can be recovered from draws of $\mathbf{B}_0, \mathbf{B}_1, \ldots, \mathbf{B}_p$, $\mathbf{\Phi}_0, \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p$, $\mathbf{\Omega}$, and $\mathbf{\Lambda}$ using the procedure provided in Chan and Eisenstat (2014).

A particular identification scheme can be imposed through a dogmatic prior which sets probability one to a particular value for κ (e.g. allocating prior probability one to

[1] These priors are noninformative in the sense that the value 0.5 implies a restriction is, a priori, equally likely to apply as not. Other priors can easily be accommodated.

$\kappa_1 = \cdots = \kappa_n = p$ will be equivalent to estimating an unrestricted VARMA(p, p)). In this case, we can work directly with $\boldsymbol{\gamma}^E$ (i.e. instead of $\boldsymbol{\gamma}^R$) to enforce the echelon form restrictions, and the SSVS indicators $\boldsymbol{\gamma}^S = \{\gamma_{ijk}^{B,S}, \gamma_{ijk}^{\Phi,S}\}$ would then be used exclusively to control additional shrinkage: they can either be fixed *a priori* with $\mathbb{P}(\gamma_{ijk}^{\cdot,S} = 1) = 1$ such that no additional shrinkage/variable selection is employed, or specified as $\mathbb{P}(\gamma_{ijk}^{\cdot,S} = 1) = 0.5$ and sampled in the course of the MCMC run along with the other parameters. Applying the latter and naively setting $\kappa_1 = \cdots = \kappa_n = p$ leads to a simple SSVS model where the parameters are potentially unidentified, but parsimony is achieved through shrinkage and computation is very fast.

Working with stochastic $\boldsymbol{\kappa}$ through stochastic row degrees $p_1, \ldots, p_n$ and indicators $\gamma_{ijk}^{B,S}$, $\gamma_{ijk}^{\Phi,S}$ as outlined above, on the other hand, results in an algorithm that always operates on a parameter space restricted according the echelon form, but also allows for additional shrinkage on the unrestricted coefficients. One interesting consequence of this is that, unlike the classic VARMA$_E(\boldsymbol{\kappa})$ model in which the number of AR coefficients must equal the number of MA coefficients, the additional SSVS priors allows the stochastic search algorithm to uncover a VARMA(p, q) where $p \neq q$ (i.e. if the SSVS mechanism additionally forces certain coefficients to zero).

In sum, we argue that this SSVS prior can successfully address two of the three reasons (identification and parsimony) for a dearth of empirical work which uses VARMAs outlined in the introduction. The third reason was computation. Our MCMC algorithm, described in Appendix B, is fairly efficient and we have had success using it in quite large VARMAs. For instance, we present empirical work below for VARMAs with $n = 12$ which is much larger than anything we have found in the existing literature with the exception of Dias and Kapetanios (2013). However, dealing with much higher dimensional models (e.g. $n = 25$ or more) as has been sometimes done with VARs would represent a serious, possibly insurmountable computational burden, with our algorithm. Furthermore, real time forecasting exercises, requiring repeated MCMC estimation on an expanding window of data, would pose challenges to our algorithm even with $n = 12$.

For these reasons, in Appendix B, we also describe an approximate MCMC algorithm which is much faster. This latter algorithm is achieved by replacing (7), which involves prior dependencies between restrictions, with the simpler independent choice $\Pr\left(\gamma_{ijk}^{B,S} = 1\right) = 0.5$. In our artificial data experiments (see below), this approximate algorithm (which we call the "row degree algorithm") seems to work quite well and is much more efficient than our exact algorithm (which we call the "echelon algorithm"). Complete details are given in Appendix B, but to understand the intuition underlying the approximate algorithm observe that (7) creates cross-equation relationships among indicators, and therefore, strong dependence between the row degrees $p_1, \ldots, p_n$. For MCMC, this forces us to sample each p_i conditional on all other row degrees and keep track of all these relationships.

However, simplifying the prior on $\gamma_{ijk}^{B,S}$ as provided above allows the approximate row degree algorithm to just draw from the row degrees ignoring the other restrictions implied by the echelon form. In this case, the row degrees are conditionally independent of one another and the MCMC algorithm can become much more efficient. This algorithm has the drawback of ignoring some restrictions of the echelon form. But this drawback, in

practice, may be slight since the SSVS prior on the VARMA coefficients (i.e. involving $\boldsymbol{\gamma}^S$) should be able to pick up any restrictions missed by using an approximate algorithm. Thus, the row degree algorithm may be useful for the researcher who finds our echelon form algorithm too computationally demanding.

The MCMC algorithms described above can be used for selecting identifying restrictions or deciding whether individual coefficients are zero or not. Of course, alternative methods of model comparison, involving marginal likelihoods or information criteria can be done using MCMC output. In our empirical section, we use the Deviance Information Criterion (DIC) for model comparison. Appendix C includes more details, including a definition and explanation of how we calculate it.

3.3 Extensions

In our empirical work, we use the models described in the preceding sub-section. However, we note that many extensions are possible. In this sub-section, we describe two directions which may be of use for the empirical macroeconomist. The first is to allow for a time-varying $\boldsymbol{\Omega}_t$ or $\boldsymbol{\Sigma}_t$. This can be done in a standard way by adding appropriate blocks to the MCMC algorithm. For instance, multivariate stochastic volatility of the form used in Primiceri (2005) can be included by adding the extra blocks to the MCMC algorithm as described in Appendix A of his paper.

A second extension we consider is related to an alternative approach to analyzing medium and large datasets. Specifically, let $\mathbf{y}_t$ be an $n \times 1$ vector of dependent variables that is categorized as follows:

- $\mathbf{y}_{1,t}$: the n_1 variables of primary interest;
- $\mathbf{y}_{2,t}$: the n_2 variables that together with $\mathbf{y}_t$ constitute a full n_1+n_2 variate VARMA process;
- $\mathbf{y}_{3,t}$: the n_3 additional variables that are used to identify factors $\mathbf{f}_t$.

Then, consider the following expanded form representation of the VARMA model:

$$\mathbf{y}_t = \sum_{j=1}^{p} \mathbf{B}_j \mathbf{y}_{t-j} + \sum_{j=0}^{q} \boldsymbol{\Phi}_j \mathbf{f}_{t-j} + \boldsymbol{\eta}_t, \qquad \mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_t) \text{ and } \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}), \tag{8}$$

where $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_q$ are $n \times n_1$ coefficient matrices and $\mathbf{f}_t$ is $n_1 \times 1$. Consequently, the covariance matrix $\boldsymbol{\Lambda}$ is of dimension $n \times n$, whereas the time-varying covariance matrix $\boldsymbol{\Omega}_t$ is diagonal with diagonal elements $\exp(h_{1,t}), \ldots, \exp(h_{n_1,t})$, where the log-volatilities follow a random walk process.

When $n \gg n_1$, (8) becomes a dynamic factor model, or under certain restrictions, a factor-augmented vector autoregression (FAVAR). In this case, identification is achieved without the need for echelon form restrictions. However, the SSVS prior for the VARMA coefficients can be maintained to ensure parsimony.

4 Empirical Results

4.1 Artificial Data

In this section, we carry out a brief exercise with artificial data to investigate the performance of our algorithm. We focus on the identification issue and present results relating to $\boldsymbol{\kappa}$ for various versions of our algorithm. All the results in this section involve drawing 10 artificial data sets, each of $T = 100$ observations. Each data set is normalized to have mean zero and unit standard deviation. For each data set, 11,000 MCMC draws are taken and the first 1,000 of these discarded. All results are based on the benchmark prior described in Appendix A. We present results for the algorithm which imposes the echelon form exactly (labelled "echelon" in the tables below) versus the approximate algorithm which works with the row degrees (labelled "row degree" in the tables below). We also investigate the difference between the two different implementations of SSVS methods (labelled "hard SSVS" and "soft SSVS" in the tables) discussed in Section 2.

The first set of artificial data exercises uses bivariate VARMAs based on (3). Our first data generating process (DGP) is a standard identified VARMA(1,1) with $\kappa_1 = \kappa_2 = 1$. The second DGP is also a VARMA(1,1) but with $\kappa_1 = 1, \kappa_2 = 0$. In both cases, our estimating model is an VARMA$(4, 4)$. The starting value for $\boldsymbol{\kappa}$ in the MCMC algorithm is, throughout this section, always set so as to choose the VARMA(4,4). We are interested in investigating whether our algorithm can, in the context of a greatly over-parameterized model, uncover the parsimonious identified model in each case. Precise values of the parameters used in the DGPs are:

DGP1: $\beta_{11} = 0.7, \beta_{21} = 0.4, \beta_{12} = 0.2, \beta_{22} = 0.5, \theta_{11} = 0.1, \theta_{12} = 0, \theta_{21} = 0.5, \theta_{22} = 0.1, \boldsymbol{\Sigma} = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.1 \end{bmatrix}$.

DGP2: $\beta_{11} = 0.7, \beta_{21} = 0, \beta_{12} = 0.2, \beta_{22} = 0, \theta_{11} = 0.1, \theta_{12} = 0, \theta_{21} = 0, \theta_{22} = 0.0, \boldsymbol{\Sigma} = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.1 \end{bmatrix}$.

Table 1 presents summary statistics of the various estimates of $\boldsymbol{\kappa}$ for the two DGPs. It can be seen that, despite working with the over-parameterized VARMA(4,4), our algorithm is accurately choosing the identified VARMA$_E(1, 1)$ and VARMA$_E(1, 0)$ for DGP$_1$ and DGP$_2$, respectively. The cross-data-set averages of $\boldsymbol{\kappa}$ do tend to be slightly above the true values used in the DGP. In the case of κ_2 in DGP$_2$, this is of necessity (since the true value of $\kappa_2 = 0$ and κ_2 cannot be negative). For other cases, this is likely due to excessively large lag length used in the estimating model. With regards to the different variants of our algorithms, there seems little difference. In particular, the approximate row degree algorithm is yielding results which are very similar to the exact algorithm which imposes the echelon form at every draw. Overall, though, the results indicate that our algorithms are working well in identifying small VARMAs. The estimates of the parameters (not reported here) are similar to the true values used in the DGPs and the inefficiency factors for the MCMC algorithm (also not reported here) indicate the algorithms are mixing well.

But will our algorithms be as capable of uncovering identification restrictions in larger VARMAs? And will they be computationally efficient? These are the questions we address

Table 1: Averages across Data Sets of Posterior Mean of $\boldsymbol{\kappa}$. Standard Deviation, Minimum and Maximum in Parentheses.

	DGP_1		DGP_2	
Algorithm details	κ_1	κ_2	κ_3	κ_4
True value	1	1	1	0
echelon, hard SSVS	1.35 (0.19) (1.14, 1.68)	1.11 (0.11) (1.02, 1.40)	1.29 (0.06) (1.21, 1.36)	0.48 (0.21) (0.31, 0.99)
row degree, hard SSVS	1.28 (0.13) (1.18, 1.63)	1.24 (0.15) (1.14, 1.63)	1.59 (0.38) (1.31, 2.54)	0.49 (0.21) (0.32, 0.91)
echelon, soft SSVS	1.28 (0.25) (1.13, 1.93)	1.09 (0.05) (1.02, 1.17)	1.30 (0.21) (1.11, 1.85)	0.49 (0.25) (0.23, 1.05)
row degree, soft SSVS	1.23 (0.13) (1.12, 1.47)	1.20 (0.14) (1.12, 1.59)	1.32 (0.21) (1.14, 1.86)	0.42 (0.30) (0.25, 1.15)

in Tables 2 through 5. Tables 2 and 4 contain results relating to $\boldsymbol{\kappa}$ comparable to those in Table 1 for larger 7-variate and 12-variate VARMAs. Tables 3 and 5 contain results relating to the efficiency of the MCMC algorithm. For the sake of brevity, inefficiency factors are presented for the impulse responses of the first and second variables to a shock in the third variable four periods in the future. These are labelled "IR_1" and "IR_2" in Tables 3 and 5.

The data generating process for the 7-variate VAR is a VARMA(1,1) with the following parameter values:

DGP_3: $\mathbf{B}_0 = \boldsymbol{\Theta}_0 = \mathbf{I}$ and $\beta_{ii} = 0.1 \times i$, $\theta_{ii} = 0.1 \times (7 - i)$, $\beta_{12} = \beta_{23} = -0.4$, $\theta_{56} = \theta_{67} = -0.4$ where β_{ij} and θ_{ij} are the (i, j) elements of $\mathbf{B}_1$ and $\boldsymbol{\Theta}_1$, respectively. Letting σ_{ij} be the elements of $\boldsymbol{\Sigma}$, we set $\sigma_{ii} = 0.1 \times i$, $\sigma_{57} = \sigma_{67} = -0.3$. All elements of $\mathbf{B}_1$, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Sigma}$ not specified are set to zero.

Note that DGP_3 has $\kappa_i = 1$ for $i = 1, \ldots, 7$ and should be well-identified in the sense that each row of the VARMA has either an AR or an MA coefficient which is substantively different from zero.

The data generating process from the 12-variate VAR is also a VARMA(1,1) with parameter values:

DGP_4: $\mathbf{B}_0 = \boldsymbol{\Theta}_0 = \mathbf{I}$ and $\beta_{ii} = 0.1 \times i$ for $i = 1, \ldots, 8$, $\theta_{ii} = 0.1 \times (12 - i)$ for $i = 1, \ldots, 10$, $\beta_{12} = \beta_{23} = -0.4$, $\theta_{56} = \theta_{67} = -0.4$ where β_{ij} and θ_{ij} are the (i, j) elements of $\mathbf{B}_1$ and $\boldsymbol{\Theta}_1$, respectively. Letting σ_{ij} be the elements of $\boldsymbol{\Sigma}$, we set $\sigma_{ii} = 0.1 \times i$, $\sigma_{57} = \sigma_{67} = -0.3$. All elements of $\mathbf{B}_1$, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Sigma}$ not specified are set to zero.

Note that DGP_4 has $\kappa_i = 1$ for $i = 1, \ldots, 10$, with $\kappa_{11} = \kappa_{12} = 0$. However, for equations 9 and 10 in the VARMA the identification is quite weak in the sense that both of these equations have no AR lags and the coefficient on the MA lag is quite small (i.e. $\theta_{99} = 0.3$ and $\theta_{10,10} = 0.2$). Hence, even through the true value $\kappa_9 = \kappa_{10} = 1$, the DGP

is quite close to the $\kappa_9 = \kappa_{10} = 0$ case.

Results for the medium-sized 7-variate VARMA are similar to those for the bivariate VARMA. Table 2 indicates the variants of our algorithm are all successfully producing an estimate of $\boldsymbol{\kappa}$ near its true value. For none of the data sets do any of our algorithms go far wrong. Table 3 indicates that the efficiency of our algorithm is fairly good, producing inefficiency factors that are around 10 or 20. However, the inefficiency factors for the echelon form algorithm with hard SSVS are somewhat higher than this. One of the artificial data sets leads to an inefficiency factor of over 300 for one of the impulse responses. Hence, the researcher using our algorithm in VARMAs of this size should take care with MCMC convergence issues and would probably be required to take hundreds of thousands of draws,[2] but MCMC convergence is unlikely to be a major worry. Indeed even the 10,000 draws (plus 1000 burn-in draws) used to produce the results in Table 2 appear to be enough to produce an accurate estimate of the true DGP in our artificial data exercise, despite the fact that the initial conditions used in our MCMC algorithm (based on the VARMA(4,4)) are far from the true VARMA(1,1).

Table 2: Averages across Data Sets of Posterior Pean of $\boldsymbol{\kappa}$ for DGP_3. Standard Deviations in Parentheses.

Algorithm details	κ_1	κ_2	κ_3	κ_4	κ_5	κ_6	κ_7
True value	1	1	1	1	1	1	1
echelon,	1.05	1.01	1.01	1.04	1.07	1.00	0.90
hard SSVS	(0.10)	(0.01)	(0.01)	(0.09)	(0.13)	(0.01)	(0.32)
row degree,	1.07	1.04	1.02	1.05	1.03	1.03	0.80
hard SSVS	(0.10)	(0.05)	(0.03)	(0.13)	(0.03)	(0.04)	(0.34)
echelon,	1.01	1.02	1.01	0.99	1.03	1.01	0.80
soft SSVS	(0.02)	(0.05)	(0.01)	(0.08)	(0.06)	(0.01)	(0.42)
row degree,	1.04	1.04	1.00	0.99	1.04	1.02	0.73
soft SSVS	(0.05)	(0.06)	(0.03)	(0.08)	(0.05)	(0.02)	(0.43)

Table 3: Inefficiency Factors for Impulse Responses for DGP_3.

Algorithm details	IR_1 ave	IR_1 st dev	IR_1 max	IR_2 ave	IR_2 st dev	IR_2 max
echelon, hard SSVS	56.38	100.12	339.57	20.71	13.74	53.76
row degree, hard SSVS	23.34	6.82	38.31	20.29	7.50	30.07
echelon, soft SSVS	13.95	6.15	25.35	15.68	9.27	33.57
row degree, soft SSVS	13.31	5.74	25.09	12.55	4.11	22.41

Results for the 12-variate VARMA are also quite encouraging. In Table 4, the estimates for $\kappa_1, \ldots, \kappa_n$ are almost always very close to the true values in the DGP. The only exception is for κ_9 and κ_{10}. But for the reasons noted previously, these are not

[2]This statement and others which follow are based on the premise that 10,000 independent draws from a posterior would produce estimates with sufficient accuracy for the researcher's purposes.

surprising. The four variants of the algorithm are producing similar results, although it is worth noting that the approximate row degree algorithms are producing estimates for κ_7 which are somewhat below those for the exact echelon algorithms.

Table 5 presents evidence on MCMC efficiency. As expected, MCMC efficiency deteriorates somewhat in this larger VARMA, but the row degree algorithm mixes much better than the echelon form algorithm. Of course, an exact algorithm is always to be preferred to an approximate one and, hence, where computationally possible we would recommend using the echelon form algorithm. However, Table 5 indicates that in larger VARMAs, the echelon form algorithm might be excessively computationally daunting or even infeasible in a reasonable amount of time. For instance, when using the echelon form algorithm with hard SSVS, one of our artificial data sets produces an inefficiency factor of over 1000 for estimation of one of the impulse responses suggesting that millions of draws may be required in some applications with larger VARMAs. In such applications, our approximate row degree algorithm, which is quite efficient even in the 12-variate VARMA, may be a good alternative.

It is also worth noting that MCMC algorithms using soft SSVS are much more efficient than hard SSVS. Even in the 12-variate VARMA, the echelon form algorithm with soft SSVS is producing inefficiency factors that are consistent with the researcher using tens (or at most a few hundred) of thousands of draws.

Table 4: Averages across Data Sets of Posterior Mean of $\boldsymbol{\kappa}$ for DGP_4. Standard Deviations in Parentheses.

Algorithm details	κ_1	κ_2	κ_3	κ_4	κ_5	κ_6
True value	1	1	1	1	1	1
echelon,	1.05	1.09	0.91	0.99	1.24	1.24
hard SSVS	(0.11)	(0.18)	(0.23)	(0.01)	(0.31)	(0.41)
row degree,	1.02	1.00	0.71	0.98	0.94	1.00
hard SSVS	(0.06)	(0.01)	(0.42)	(0.04)	(0.23)	(0.00)
echelon,	0.99	1.00	0.79	0.95	1.18	1.02
soft SSVS	(0.01)	(0.00)	(0.36)	(0.14)	(0.39)	(0.03)
row degree,	0.99	1.00	0.67	0.96	0.92	1.00
soft SSVS	(0.01)	(0.00)	(0.44)	(0.11)	(0.23)	(0.00)
	κ_7	κ_8	κ_9	κ_{10}	κ_{11}	κ_{12}
True value	1	1	1	1	0	0
echelon,	1.00	0.96	0.04	0.02	0.00	0.00
hard SSVS	(0.00)	(0.12)	(0.01)	(0.03)	(0.00)	(0.00)
row degree,	0.33	0.98	0.03	0.02	0.00	0.00
hard SSVS	(0.39)	(0.05)	(0.06)	(0.02)	(0.00)	(0.00)
echelon,	0.81	0.93	0.01	0.01	0.01	0.00
soft SSVS	(0.41)	(0.16)	(0.01)	(0.01)	(0.00)	(0.00)
row degree,	0.21	0.94	0.01	0.01	0.00	0.00
soft SSVS	(0.39)	(0.16)	(0.02)	(0.02)	(0.00)	(0.00)

Table 5: Inefficiency Factors for Impulse Responses for DGP_4.

Algorithm details	IR_1 ave	IR_1 st dev	IR_1 max	IR_2 ave	IR_2 st dev	IR_2 max
echelon, hard SSVS	152.98	274.74	766.9	234.73	411.56	1169.9
row degree, hard SSVS	17.44	9.49	28.57	24.31	25.77	88.91
echelon, soft SSVS	36.02	58.27	179.48	31.26	54.74	178.57
row degree, soft SSVS	10.12	3.13	16.80	10.54	4.51	20.42

In summary, our artificial data exercise shows that our approach performs well in picking out the correct restrictions required to choose a correctly-identified parsimonious VARMA in the context of estimating an unidentified over-parameterized VARMA(4,4) even when n is quite large. The findings also lead us to recommend the use of soft SSVS over hard SSVS for MCMC efficiency reasons and our empirical results using macroeconomic data will use soft SSVS. We also recommend the use of our exact echelon form algorithm as opposed to the approximate row degree algorithm where possible. But we include the row degree algorithm in this paper since in larger VARMAs it may be required. Our empirical results using macroeconomic data use the echelon form algorithm.

4.2 Macroeconomic Application

In this section, we investigate the performance of our echelon algorithm (using soft SSVS and the prior specified in Appendix A) in a substantive empirical application involving quarterly US macroeconomic data in VARMAs of varying dimensions: $n = 3$, $n = 7$ and $n = 12$. We will draw all inference from a VARMA$(4, 4)$, of the form

$$\mathbf{y}_t = \sum_{j=1}^{4} \mathbf{A}_j \mathbf{y}_{t-j} + \sum_{j=1}^{4} \mathbf{\Theta}_j \boldsymbol{\epsilon}_{t-j} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}). \tag{9}$$

For quarterly data, we conjecture that this is potentially over-parameterized. Therefore, for estimation purposes we will employ the echelon form and rely on the data to uncover the correct Kronecker structure. Our MCMC algorithms obtain draws directly from the expanded form. Upon the termination of the MCMC routine, however, we transform all draws ex post to recover $\mathbf{A}_1, \ldots, \mathbf{A}_4$, $\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_4$, $\mathbf{\Sigma}$ in (9) above. We then analyze estimates of these parameters and compute impulse responses. In sub-section 4.2.2 we also consider a simpler approach that involves sampling from (9) without employing the echelon form, but only specifying SSVS shrinkage priors on the over-parameterized VARMA$(4, 4)$. We compare the results obtained with both approaches, as well as those generated by a VAR(4) with SSVS priors on the VAR coefficients.

Our data covers the quarters 1959:Q1 to 2013:Q4. As is commonly done (e.g., Stock and Watson, 2008) and recommended in Carriero, Clark and Marcellino (2011), each series is transformed to stationarity. We use a recursive identification scheme for our impulse responses following standard practice when working with large macroeconomic data sets (e.g. Bernanke, Boivin, and Eliasz, 2005, and Banbura, Giannone and Reichlin, 2010). In particular, we treat the Federal Funds rate as the monetary policy instrument (which is

orthogonal to all other shocks) and classify every other variable as either "slow-moving" or "fast-moving" relative to this. Variables are ordered as slow-moving, then the monetary policy instrument, then the fast-moving variables. We stress that our variables have been transformed (e.g. GDP is log differenced) and that impulse responses reported below are to these transformed variables. Exact definitions of the variables, their transformations and classifications are given in Appendix D.

4.2.1 Results for our Preferred Model

In this sub-section, we focus on our preferred approach, as described in the preceding sub-section. We run the algorithm for 50,000 iterations (5,000 burn-in) for the $n = 3$ model, 200,000 iterations (20,000 burn-in) for the $n = 7$ model, and 1,000,000 iterations (100,000 burn-in) for the $n = 12$ model. For each model, we then thin the chains to obtain exactly 10,000 draws (e.g., for $n = 3$ we take every 5th draw, for $n = 7$ every 20th draw and for $n = 12$ every 100th draw). In each case, we set $\kappa_{\max} = 4$.

Figure 1 presents the estimated impulses responses of GDP, inflation and the interest rate to a shock in the interest rate, for 20 quarters following the shock. Table 6 presents inefficiency factors relating to these impulse responses. Specifically, it contains summary statistics for the inefficiency factors of the 60 different impulse responses computed and indicates that the number of draws taken is longer than necessary if one is only interested in obtaining impulse responses.

Table 6: Comparison of inefficiency factors for impulse responses across the three models: $n = 3$, $n = 7$, and $n = 12$; note that the reported inefficiency factors are computed on thinned draws.

n	IF avg	IF st dev	IF max
3	5.90	3.17	16.10
7	1.86	2.38	15.22
12	1.17	0.43	2.90

Since we are interested in accurately estimating the Kronecker indices, we also present results on MCMC performance relating to them. However, since $\kappa_1, \ldots, \kappa_n$ are discrete random variables, inefficiency factors are not an appropriate way to gauge sampling efficiency. In addition, any particular κ_i may naturally exhibit little movement over the course of the sampler. For instance, if there is one correct choice for κ_i then a good MCMC sampler would often (or even always) make such a choice and a lack of switching in the chain could be consistent with good MCMC performance. Accordingly, we shed light on the efficiency of the algorithm by the number of times the sampler switches models, as defined by the entire vector $\boldsymbol{\kappa}$. Specifically, we compute the metric

$$\varpi_n = \sum_{g=1}^{G} \mathbb{1}\left(\sum_{i=1}^{n} \left| \kappa_i^{(g)} - \kappa_i^{(g-1)} \right| > 0 \right) / G,$$

where G is the number of MCMC draws, and consider that 10% represents sufficient mobility for estimation purposes. This metric is reported in Table 7 along with the

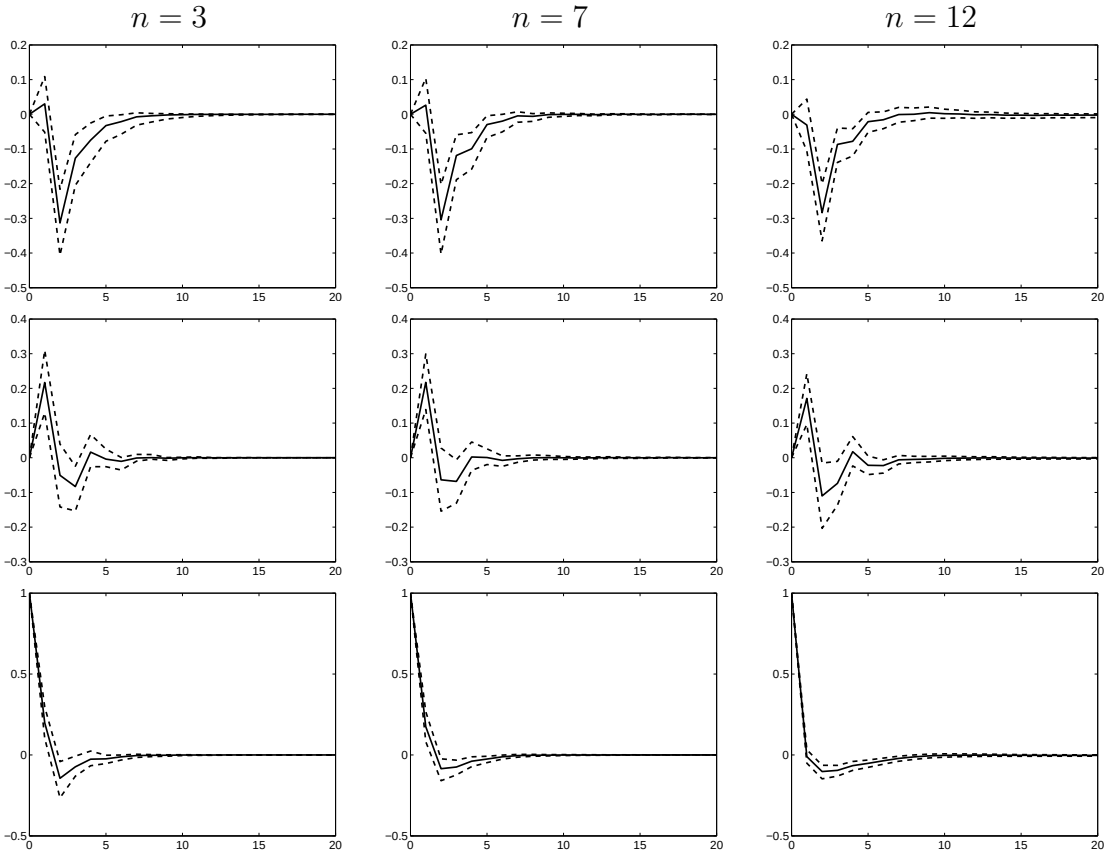

Figure 1: Impulse responses to a shock in the interest rate. The first row contains responses of GDP to a shock in the interest rate; the second row contains responses of inflation to a shock in the interest rate; the third row contains responses the interest rate to its own shock. The dotted lines depict the $(10\%, 90\%)$ HPD intervals.

estimated $\boldsymbol{\kappa}$ for the VARMAs of different dimensions. Two general points are worth noting: the MCMC sampler is mixing well and the identification restrictions selected are much more parsimonious than the VARMA(4,4) estimating model. These facts suggest our modelling approach and associated MCMC algorithm are working well, even in large VARMAs.

A specific point worth noting is that, for output and inflation, the estimated Kronecker indices are consistent across VARMAs of different dimensions. In contrast, the Kronecker index for the interest rate decreases as the size of the system increases. This result is related to the ordering of the variables and is, in fact, consistent with the Kronecker index theory. Loosely speaking, a Kronecker index κ_i represents a threshold beyond which autocovariances of further lags are linearly dependent on the lower-degree autocovariances of variables $1, \ldots, i$. Since output and inflation are always ordered first, we expect that the associated Kronecker indices do not change as additional variables are introduced. However, moving from three variables to seven, and especially from seven to twelve, introduces new variables that precede the interest rate. The fact that the Kronecker index on the interest rate shrinks from an estimated $\hat{\kappa}_3 = 1.68$ for the $n = 3$ system to $\hat{\kappa}_9 = 0.01$ for the $n = 12$ system indicates that the additional variables contain all necessary information to explain the autocorrelations present in the interest rate. In other words, we infer from the $n = 12$ system that the interest rate only responds contemporaneously to slow moving variables; removing these variables from the model leads us to estimate the interest rate as an autocorrelated process.

Table 7: Comparison of estimated Kronecker indices across the three models: $n = 3$, $n = 7$, and $n = 12$

		n = 3	n = 7	n = 12
1	Real Gross Domestic Product	2.05	1.99	2.00
2	Consumer Price Index: All Items	2.12	2.01	2.00
3	Real Personal Consumption Exp.			1.00
4	Housing Starts: Total			1.00
5	Average Hourly Earnings: Manuf.		3.00	3.00
6	Real Gross Private Domestic Invest.			1.00
7	All Employees: Total nonfarm			1.00
8	ISM Manuf.: PMI Composite Index			1.00
9	Effective Federal Funds Rate	1.68	0.99	0.01
10	S&P 500 Stock Price Index		1.00	0.84
11	M2 Money Stock		1.28	0.97
12	Spot Oil Price: West Texas Interm		0.88	0.40
ϖ_n		23.8%	13.7%	13.4%

The preceding table suggests our methodology is successfully picking out parsimonious identified models. To investigate this issue more deeply, Tables 8-9 present estimates of the autoregressive and moving average coefficients in the $n = 12$ model. For comparison, we also present the estimates of autoregressive coefficients obtained from a 12-variate VAR(4) in Table 10. The VAR is obtained by starting with the echelon form VARMA,

discarding the echelon restrictions, and setting $q = 0$. We then use the algorithm described in Section 2.

In Tables 8-9 it can be seen that the matrices of AR and MA coefficients are mostly zeros, particularly at longer lag lengths. This strengthens the evidence in support of our specification and algorithm successfully achieving parsimony. Note also that there are several non-zero coefficients in $\mathbf{\Theta}_1$ (and some in $\mathbf{\Theta}_2$) indicating that adding MA terms to the VAR is important. A careful examination of the MA coefficients shows that it is usually errors in the housing starts and the purchasing manager's index equations that are found to be important. It is interesting to note that these two variables are typically regarded as leading indicators. Results for the housing starts variable are particularly interesting. When estimating the VARMA, we are finding in most equations that housing starts' effect is best modelled through the MA part of the model. That is, other variables typically react to innovations in the housing starts equation, not lags of the housing starts variable itself (i.e. the fourth columns of $\mathbf{A}_1, \ldots, \mathbf{A}_4$ are mostly zeros). Of course, the VAR itself could not produce such a finding. It is interesting to note in Table 10 that lagged housing starts now appear much more prominently in the VAR part of the model, included in some equations at the second or third lag. This is as theory would predict. A parsimonious VARMA, such as that obtained in Tables 8-9, may be approximated by a VAR. However, the resulting VAR will be less parsimonious and with a longer lag length. In other words, it looks like the VAR(4) is trying to fit an inverted VARMA process.

4.2.2 Comparison with Alternative Approaches

In order to investigate the advantages of working with a VARMA over a VAR and the importance of imposing identification, in this sub-section we compare our preferred approach to a different VARMA (which does have prior shrinkage but does not explicitly impose identification) and a VAR (which does have shrinkage but no MA components). In particular, for each model of dimension n, we compare the following specifications:

- VARMA$_E(\boldsymbol{\kappa})$: our preferred echelon form VARMA with soft SSVS priors on AR and MA coefficients and $\kappa_{\max} = 4$;
- VARMA$(4, 4)$: a VARMA with soft SSVS priors but no echelon form restrictions;
- VAR(4): a VAR with soft SSVS priors.

Note that we are only comparing modelling approaches which involve prior shrinkage. As we shall see below, empirical results such as impulse responses are clearly inferior and imprecise when we do not do have such shrinkage.

We begin by calculating DICs (see Appendix C) for each model and report the results in Table 11. It can be seen that, for models of larger dimension, our VARMA$_E(\boldsymbol{\kappa})$ is the preferred model by a substantial margin. Although when $n = 3$ the VARMA$(4, 4)$ is preferred.

Each column in Table 11 contains results for a different value of n and, thus, a different y_t. Hence, results are not comparable across columns and the table cannot be used to provide evidence for or against working with a large dimensional model. Table 12 repeats

Table 8: Posterior estimates of the moving average coefficients matrices $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_4$ in a $\text{VARMA}_E(\boldsymbol{\kappa})$. Note: * denotes that either $\Pr(\theta_{l,ij} \leq 0 \,|\, \mathbf{y}) \leq 0.1$ or $\Pr(\theta_{l,ij} > 0 \,|\, \mathbf{y}) \leq 0.1$; § denotes that either $\Pr(\theta_{l,ij} \leq 0 \,|\, \mathbf{y}) \leq 0.05$ or $\Pr(\theta_{l,ij} > 0 \,|\, \mathbf{y}) \leq 0.05$; † denotes that either $\Pr(\theta_{l,ij} \leq 0 \,|\, \mathbf{y}) \leq 0.01$ or $\Pr(\theta_{l,ij} > 0 \,|\, \mathbf{y}) \leq 0.01$.

		1	2	3	4	5	6	7	8	9	10	11	12
	1	-0.19*	-0.01	0.06	0.47§	-0.01	-0.04	0.20*	0.12	-0.02	0.06	0.00	-0.03
	2	-0.06	-0.08	0.07	0.03	0.01	-0.12	0.05	0.23§	0.03	0.01	0.01	0.04
	3	-0.09	-0.15	0.02	0.53§	0.00	0.04	-0.07	-0.12	-0.08	0.04	-0.03	-0.07
	4	-0.02	-0.01	0.01	0.08§	0.00	0.00	0.01	-0.01	-0.01	0.01	0.00	-0.01
	5	0.00	0.04	-0.05	0.17	0.01	-0.03	-0.03	0.05	-0.01	0.01	0.00	-0.01
$\boldsymbol{\Theta}_1$	6	-0.20*	0.06	0.01	0.66†	-0.03	-0.08	0.38†	0.20*	0.02	0.03	0.03	0.00
	7	-0.08	-0.01	0.02	0.25§	-0.02	-0.04	0.03	0.13§	-0.01	0.02	0.00	-0.01
	8	-0.07	-0.02	0.00	0.30†	-0.01	-0.03	0.06	0.11*	-0.02	0.02	0.01	-0.01
	9	-0.05	-0.02	0.01	0.15§	-0.01	-0.04	0.05	0.11§	0.00	0.01	0.01	0.00
	10	-0.03	-0.03	0.01	0.26*	0.00	0.00	-0.07	-0.01	-0.04	0.06	0.01	-0.02
	11	0.07	-0.02	0.00	-0.15	0.04	0.05	-0.12	-0.26§	-0.03	0.01	-0.09	-0.01
	12	-0.01	0.01	0.03	-0.11	0.00	-0.04	0.01	0.13*	0.01	0.01	0.02	0.03
	1	-0.03	-0.03	-0.01	0.11	-0.01	-0.02	0.12	0.02	-0.04	-0.01	0.00	-0.01
	2	0.04	-0.14	-0.02	-0.04	0.04	-0.01	-0.05	-0.11	-0.02	-0.02	-0.03	-0.09
	3	-0.02	-0.02	0.00	0.07	-0.01	-0.01	0.08	0.01	-0.03	-0.01	0.00	-0.01
	4	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
	5	0.12	-0.05	-0.13*	-0.01	-0.04	-0.11	0.24§	0.04	0.00	-0.01	0.00	0.00
$\boldsymbol{\Theta}_2$	6	-0.01	-0.02	-0.01	0.07	-0.01	-0.02	0.10	0.02	-0.03	-0.01	0.00	-0.01
	7	0.00	-0.02	-0.01	0.03	0.00	-0.02	0.05	0.00	-0.01	-0.01	0.00	-0.01
	8	0.01	-0.02	-0.02	0.02	-0.01	-0.02	0.06*	0.00	-0.01	0.00	0.00	-0.01
	9	0.01	-0.03	-0.01	0.01	0.00	-0.01	0.01	-0.02	-0.01	-0.01	-0.01	-0.02
	10	-0.02	0.00	0.01	0.03	0.00	0.00	0.01	0.00	-0.01	0.00	0.00	0.00
	11	-0.02	0.07	0.01	0.01	-0.01	0.01	0.00	0.04	0.01	0.01	0.02	0.04
	12	0.02	-0.03	0.00	-0.02	0.01	0.00	-0.03	-0.04	0.00	-0.01	-0.01	-0.02
	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5	-0.02	0.00	0.02	-0.05	-0.01	0.00	-0.02	0.08	0.00	0.01	0.01	0.01
$\boldsymbol{\Theta}_3$	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	8	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00
	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{\Theta}_4$	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 9: Posterior estimates of the autoregressive coefficients matrices $\mathbf{A}_1, \ldots, \mathbf{A}_4$ in a VARMA$_E(\boldsymbol{\kappa})$. Note: * denotes that either $\Pr(\theta_{l,ij} \leq 0 \,|\, \mathbf{y}) \leq 0.1$ or $\Pr(\theta_{l,ij} > 0 \,|\, \mathbf{y}) \leq 0.1$; § denotes that either $\Pr(\theta_{l,ij} \leq 0 \,|\, \mathbf{y}) \leq 0.05$ or $\Pr(\theta_{l,ij} > 0 \,|\, \mathbf{y}) \leq 0.05$; † denotes that either $\Pr(\theta_{l,ij} \leq 0 \,|\, \mathbf{y}) \leq 0.01$ or $\Pr(\theta_{l,ij} > 0 \,|\, \mathbf{y}) \leq 0.01$.

		1	2	3	4	5	6	7	8	9	10	11	12
	1	0.01	-0.01	0.12*	0.08	0.06	0.01	0.07	0.09	-0.01	0.11*	-0.02	-0.05
	2	0.03	-0.54†	0.04	0.01	0.01	0.00	0.01	0.12*	0.15§	-0.02	0.02	0.08
	3	0.05	-0.05	0.03	0.12*	0.00	0.10	0.09	0.05	-0.15§	0.12*	-0.07	0.00
	4	-0.04	0.03	0.04*	0.97†	-0.01	0.06§	0.01	-0.05§	-0.07†	0.04*	0.01	0.00
	5	-0.01	0.03	-0.06	0.05	-0.80†	-0.03	0.03	0.03	-0.02	0.03	0.01	0.02
$\mathbf{A}_1$	6	-0.08	-0.02	0.29†	-0.03	0.06	0.07	-0.07	0.12*	0.08	0.16§	-0.01	-0.03
	7	-0.02	-0.08§	0.11§	0.03	-0.02	0.02	0.67†	0.11§	-0.02	0.12†	-0.03	0.03
	8	0.11*	-0.02	0.07	0.05	-0.05*	-0.03	-0.15§	0.83†	-0.10§	0.10§	-0.07*	-0.01
	9	0.04	-0.10§	0.06*	0.05	0.00	-0.01	0.02	0.33†	-0.01	0.05§	-0.02	0.01
	10	0.06	0.05	0.00	-0.01	0.10*	0.04	0.04	-0.28§	-0.05	0.19§	-0.03	-0.02
	11	-0.09	-0.06	0.03	-0.02	-0.09	-0.10	0.04	0.03	-0.07	0.04	-0.15*	0.01
	12	0.01	-0.05	-0.03	0.05	0.00	0.00	-0.02	0.09	0.02	0.00	-0.01	0.05
	1	0.04	0.02	0.12*	0.00	-0.01	0.07	-0.06	0.01	-0.17†	0.05	0.03	-0.04
	2	0.02	-0.22§	0.09	0.00	-0.01	-0.14§	-0.02	0.05	0.02	-0.03	0.05	-0.10
	3	0.03	0.02	0.08*	0.00	-0.01	0.05	-0.04	0.00	-0.12†	0.04	0.02	-0.02
	4	0.00	0.00	0.01*	0.00	0.00	0.01	-0.01	0.00	-0.02§	0.00	0.00	-0.01
	5	0.04	0.04	-0.04	0.02	-0.70†	-0.02	0.07	0.01	-0.04	-0.03	-0.02	0.01
$\mathbf{A}_2$	6	0.03	0.03	0.08*	0.00	-0.06	0.05	-0.04	0.00	-0.12†	0.04	0.02	-0.02
	7	0.02	0.00	0.04*	0.00	-0.06§	0.01	-0.02	0.01	-0.06§	0.01	0.01	-0.02
	8	0.02	0.00	0.03	0.00	-0.09†	0.01	-0.01	0.01	-0.05§	0.01	0.01	-0.01
	9	0.01	-0.03*	0.03*	0.00	-0.03	-0.02	-0.01	0.01	-0.02	0.00	0.01	-0.02
	10	0.01	0.00	0.03*	0.00	0.05	0.02	-0.02	0.00	-0.04*	0.01	0.01	-0.01
	11	-0.01	0.08§	-0.05	0.00	0.01	0.05*	0.01	-0.02	0.01	0.01	-0.03	0.04
	12	0.00	-0.08§	0.01	0.00	0.01	-0.05*	0.00	0.02	0.03	-0.02	0.01	-0.04
	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.00	-0.03	-0.08	0.01	-0.53†	0.07	-0.01	-0.01	-0.07	0.05	0.09*	0.09*
$\mathbf{A}_3$	6	0.00	0.00	-0.01	0.00	-0.04§	0.01	0.00	0.00	-0.01	0.00	0.01*	0.01*
	7	0.00	0.00	-0.01	0.00	-0.04†	0.01	0.00	0.00	-0.01	0.00	0.01*	0.01*
	8	0.00	0.00	-0.01	0.00	-0.07†	0.01	0.00	0.00	-0.01	0.01	0.01*	0.01*
	9	0.00	0.00	0.00	0.00	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.01	0.00	0.04	-0.01	0.00	0.00	0.01	0.00	-0.01	-0.01
	11	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	12	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\mathbf{A}_4$	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 10: Posterior estimates of the autoregressive coefficients matrices $\mathbf{A}_1, \ldots, \mathbf{A}_4$ in a VAR(4). Note: * denotes that either $\Pr(\theta_{l,ij} \leq 0 \,|\, \mathbf{y}) \leq 0.1$ or $\Pr(\theta_{l,ij} > 0 \,|\, \mathbf{y}) \leq 0.1$; § denotes that either $\Pr(\theta_{l,ij} \leq 0 \,|\, \mathbf{y}) \leq 0.05$ or $\Pr(\theta_{l,ij} > 0 \,|\, \mathbf{y}) \leq 0.05$; † denotes that either $\Pr(\theta_{l,ij} \leq 0 \,|\, \mathbf{y}) \leq 0.01$ or $\Pr(\theta_{l,ij} > 0 \,|\, \mathbf{y}) \leq 0.01$.

		1	2	3	4	5	6	7	8	9	10	11	12
	1	-0.10	-0.03	0.12*	0.63†	0.05	-0.12	0.31§	0.15*	0.00	0.16†	-0.02	-0.06
	2	0.00	-0.64†	0.09	0.04	0.01	-0.10	0.01	0.31§	0.19†	0.02	0.02	0.17†
	3	-0.03	-0.20†	0.00	0.57†	-0.03	0.05	0.15	-0.02	-0.22†	0.17†	-0.06	-0.05
	4	-0.05	0.00	0.03	1.10†	-0.01	0.06*	0.02	-0.06*	-0.09†	0.05§	0.00	0.00
	5	-0.04	0.07	-0.09*	0.12	-0.79†	-0.05	0.05	0.10	-0.06	0.04	0.03	0.01
$\mathbf{A}_1$	6	-0.14*	0.03	0.20†	0.86†	0.05	-0.18§	0.28§	0.22§	0.12§	0.14†	0.01	-0.01
	7	-0.06	-0.09§	0.11§	0.37†	-0.06*	-0.05	0.63†	0.30†	-0.03	0.11†	-0.02	0.04
	8	0.03	-0.06	0.04	0.33†	-0.06	-0.12§	0.02	0.91†	-0.10§	0.14†	-0.03	0.00
	9	0.02	-0.19†	0.01	0.33§	-0.03	-0.06	0.05	0.58†	0.08	0.06	0.06	0.11§
	10	0.05	-0.03	0.01	0.30*	0.08	0.02	0.00	-0.35§	-0.04	0.29†	0.01	0.00
	11	-0.07	-0.02	0.07	-0.14	0.01	-0.05	-0.06	-0.30†	-0.16§	0.06	-0.33†	-0.07
	12	-0.01	0.02	0.02	-0.04	-0.02	-0.01	-0.01	0.22*	0.01	0.07	-0.03	0.15§
	1	-0.01	-0.02	0.07	-0.27	-0.02	-0.05	0.07	0.03	-0.16§	0.00	0.04	-0.05
	2	-0.02	-0.42†	0.05	-0.03	0.04	-0.20§	-0.04	-0.12	0.10*	-0.08*	0.05	-0.18†
	3	0.00	-0.09	0.10	-0.22	-0.06	-0.02	0.08	0.14	-0.15§	0.03	0.02	0.01
	4	-0.05	-0.04	0.05*	0.00	-0.02	0.03	-0.03	0.08*	-0.01	-0.02	-0.02	-0.01
	5	0.06	0.00	-0.11§	-0.02	-0.76†	-0.04	0.25§	0.03	-0.04	-0.03	-0.01	0.00
$\mathbf{A}_2$	6	-0.02	0.01	0.01	-0.38*	-0.06	-0.02	-0.06	-0.04	-0.04	0.02	0.04	-0.02
	7	0.05	-0.02	-0.02	-0.17	-0.08*	-0.08	0.03	-0.17§	0.02	0.05*	0.05	-0.02
	8	0.05	-0.03	-0.07*	-0.18	-0.11§	-0.11§	0.07	-0.01	0.00	-0.02	0.01	-0.04
	9	0.04	-0.06	-0.23†	-0.16	0.00	-0.09	0.25*	-0.21	-0.17§	-0.01	0.06	-0.04
	10	-0.05	-0.08	0.02	-0.19	0.06	0.00	0.04	0.01	0.11*	0.00	0.04	-0.02
	11	-0.03	0.07	0.09	0.30	0.05	0.14*	-0.09	0.27§	0.02	0.07	-0.30†	0.12§
	12	-0.10	-0.12	-0.06	0.12	0.03	-0.05	-0.01	-0.10	0.17§	-0.05	0.06	-0.13*
	1	-0.01	-0.01	0.00	-0.07	0.01	0.00	-0.31§	0.01	0.06	0.02	0.00	-0.02
	2	0.01	-0.05	-0.02	0.01	0.08	0.01	-0.07	0.13	-0.01	0.12§	-0.03	-0.10*
	3	0.00	-0.08	0.09	-0.11	0.05	0.03	-0.13	0.03	-0.02	-0.04	0.02	-0.10*
	4	-0.07*	-0.05*	0.03	-0.13*	0.01	0.03	0.02	-0.07*	0.00	0.00	-0.02	0.02
	5	-0.04	-0.05	-0.05	-0.02	-0.60†	0.07	-0.08	0.03	-0.10§	0.04	0.11§	0.09§
$\mathbf{A}_3$	6	-0.02	0.01	-0.04	-0.37*	-0.07	-0.01	-0.27§	0.07	0.06	0.02	0.04	0.05
	7	0.06	-0.01	-0.02	-0.09	-0.07*	-0.01	-0.05	0.08	-0.01	0.05	0.01	-0.04
	8	0.09	-0.07*	-0.01	0.00	-0.06	-0.06	-0.19§	0.04	-0.03	0.00	0.00	-0.03
	9	-0.04	-0.01	-0.02	-0.05	-0.01	-0.10	0.01	0.23*	0.05	0.06	0.08	-0.08
	10	0.03	-0.11*	0.11*	-0.05	-0.07	-0.05	-0.07	0.01	0.12*	-0.04	-0.05	-0.09*
	11	0.03	0.00	0.02	-0.13	0.02	0.05	0.00	-0.08	0.00	0.01	-0.09	0.04
	12	-0.01	0.04	-0.03	0.02	0.07	-0.06	-0.08	0.09	0.02	0.03	-0.05	-0.02
	1	0.07	-0.04	0.03	-0.19	0.01	-0.04	0.03	0.02	-0.05	0.01	-0.01	-0.02
	2	0.02	-0.12§	-0.03	0.03	-0.02	0.06	0.03	0.04	-0.04	-0.02	0.03	-0.09*
	3	-0.01	-0.03	-0.01	-0.10	-0.01	0.03	-0.01	0.01	-0.09	0.04	-0.05	0.03
	4	0.01	0.00	0.03	-0.01	0.03*	-0.01	0.00	0.04	0.00	-0.02	0.00	-0.02
	5	-0.02	0.01	0.00	0.00	-0.01	-0.16§	0.03	-0.04	0.05	-0.02	0.03	0.02
$\mathbf{A}_4$	6	0.07	0.00	0.14§	-0.13	-0.03	-0.02	-0.05	-0.12	0.06	0.01	0.03	-0.07*
	7	0.00	-0.02	0.05	-0.09	-0.05	-0.04	0.05	0.01	-0.04	0.01	-0.05*	-0.02
	8	-0.04	-0.02	0.12§	-0.10	-0.02	-0.04	0.00	-0.04	-0.02	0.00	-0.04	0.00
	9	-0.06	-0.02	-0.04	-0.02	-0.03	-0.03	0.00	-0.13	0.04	-0.06	0.00	-0.01
	10	0.02	-0.05	-0.02	-0.08	-0.05	-0.04	-0.01	0.08	-0.07	0.04	-0.02	-0.07
	11	-0.09	-0.03	-0.02	-0.08	0.07	-0.01	0.05	0.08	0.11*	-0.04	-0.16†	0.05
	12	0.01	-0.05	0.02	0.02	-0.02	-0.01	-0.06	0.10	0.05	0.04	0.04	-0.10*

Table 11: Estimated DIC values and associated numerical standard errors (in parentheses).

	$n = 3$	$n = 7$	$n = 12$
VARMA$_E(\boldsymbol{\kappa})$	1654.8	3738.1	4674.3
	(0.46)	(0.56)	(0.64)
VARMA(4,4)	1645.5	3748.1	4685.5
	(0.38)	(0.16)	(0.27)
VAR(4)	1654.5	3763.8	4738.9
	(0.12)	(0.08)	(0.10)

Table 12: Estimated DIC values and associated numerical standard errors (in parentheses). The DICs are computed using the marginal distribution of the three variables in the $n = 3$ case as the likelihood.

	$n = 3$	$n = 7$	$n = 12$
VARMA$_E(\boldsymbol{\kappa})$	1654.8	1682.2	1653.1
	(0.46)	(2.67)	(0.21)
VARMA(4,4)	1645.5	1707.9	1687.2
	(0.38)	(0.28)	(0.17)
VAR(4)	1654.5	1657.4	1593.5
	(0.12)	(0.13)	(0.13)

the calculation of DICs, but using in the likelihood function only the three variables common to all models. This table can be used to discuss the relative merits of models of different dimension, at least in terms of their ability to explain inflation, output growth and the interest rate. This table offers support for working with the larger $n = 12$ model. For the VARMA$_E(\boldsymbol{\kappa})$ and VAR(4), the $n = 12$ model yields the best value for the DIC. Interestingly, Table 12 finds the VAR(4) with $n = 12$ to be the overall best model. This contrasts with Table 11 where the VARMA$_E(\boldsymbol{\kappa})$ with $n = 12$ was found to be superior to the VAR(4) with $n = 12$. The explanation for this contrast is that the VARMA$_E(\boldsymbol{\kappa})$ is the superior model when interest centers on jointly explaining the 12 variables in y_t. If interest centers on explaining only the 3 variables, then the MA terms are not helpful and the VAR is best (although the remaining 9 variables do provide useful explanatory power since the VAR with $n = 12$ is chosen over VARs with $n = 3$ or 7).

Next we compare impulse responses for the different models and choices for n. Figures 2, 3 and 4 plot conventional impulse responses of our three main variables to a monetary policy shock. Our findings of the preceding sub-section indicate that the housing starts variable is found to be of particular importance and, in this section, we have evidence in favor of $n = 12$. Accordingly, in Figure 5, we plot impulse responses relating to this variable for different models for $n = 12$. The overall message of these figures, and Figure 5 in particular, is that MA components and identification can have an appreciable impact on impulse responses.

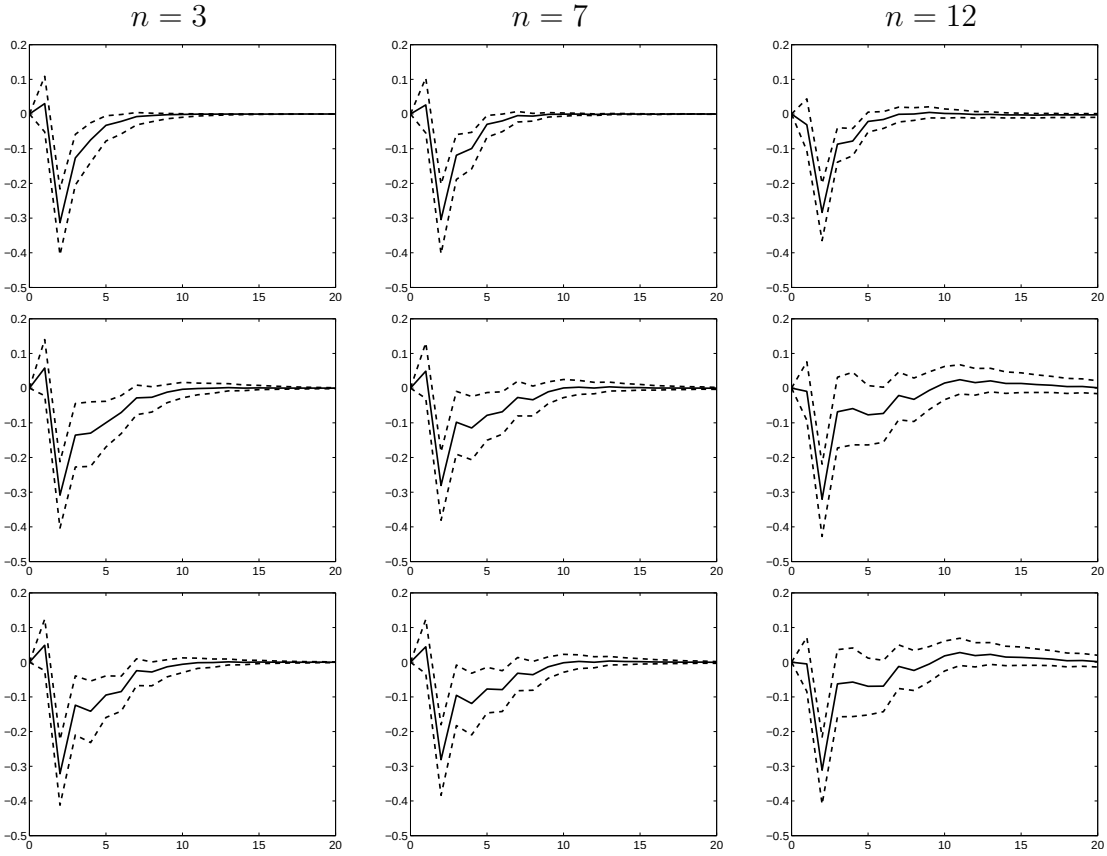

Figure 2: Comparison of impulse responses of GDP to a shock in the interest rate. The first, second and third rows contain results for the VARMA$_E(\boldsymbol{\kappa})$, VARMA$(4, 4)$ and VAR(4), respectively. The dotted lines depict the $(10\%, 90\%)$ HPD intervals.

If we compare VARMA$_E(\boldsymbol{\kappa})$, VARMA$(4, 4)$ and VAR(4) impulse responses in Figures 2, 3 and 4, we see some differences in the point estimates. The impulse responses produced by the VARMA$_E(\boldsymbol{\kappa})$ are slightly smoother, having less of the irregular up and down movements of the impulse responses produced by the other approaches, particularly for $n = 12$. Furthermore, the HPD intervals are tighter when using the VARMA$_E(\boldsymbol{\kappa})$.

In Figure 5, which plots impulse responses relating to the housing variable for the VARMA$_E(\boldsymbol{\kappa})$ and VAR(4) , we can see the role of the including MA components. The VARMA$_E(\boldsymbol{\kappa})$ is producing smooth and sensible point estimates of impulse responses with fairly tight HPD intervals about them. The VAR is producing slightly more irregular impulse responses and the HPD intervals a wider. These differences could lead to different policy conclusions. Looking at the results generated by the VARMA$_E(\boldsymbol{\kappa})$ model, it appears that the interest rate will continue to increase[3] in response to a housing starts shock, even after 20 quarters. This finding is significant in the sense that the HPD interval is entirely above zero. The housing start variable itself is very slow in adjusting downward following a positive shock, suggesting that increasing interest rates exerts little effect in discouraging further real-estate expansion. This is partly confirmed by the impulse re-

[3]Recall that the interest rate series is first-differenced.

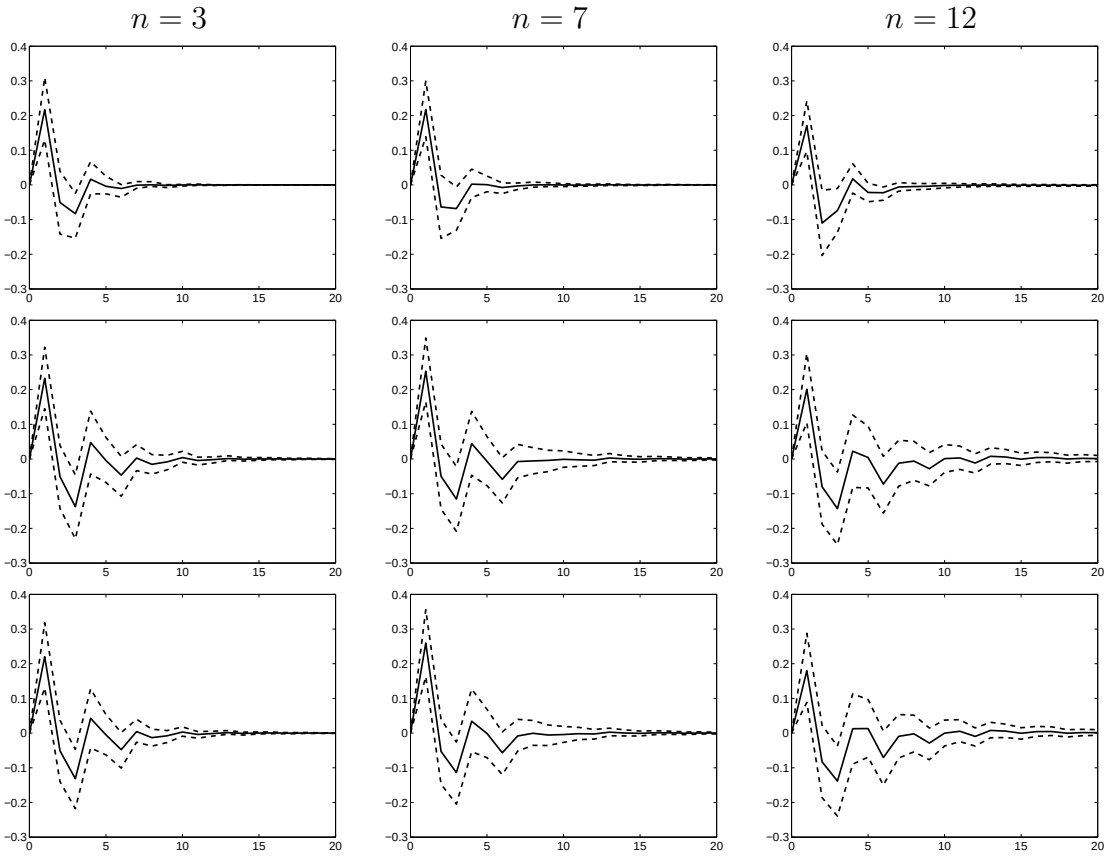

Figure 3: Comparison of impulse responses of inflation to a shock in the interest rate. The first, second and third rows contain results for the VARMA$_E(\boldsymbol{\kappa})$, VARMA$(4, 4)$ and VAR(4), respectively. The dotted lines depict the $(10\%, 90\%)$ HPD intervals.

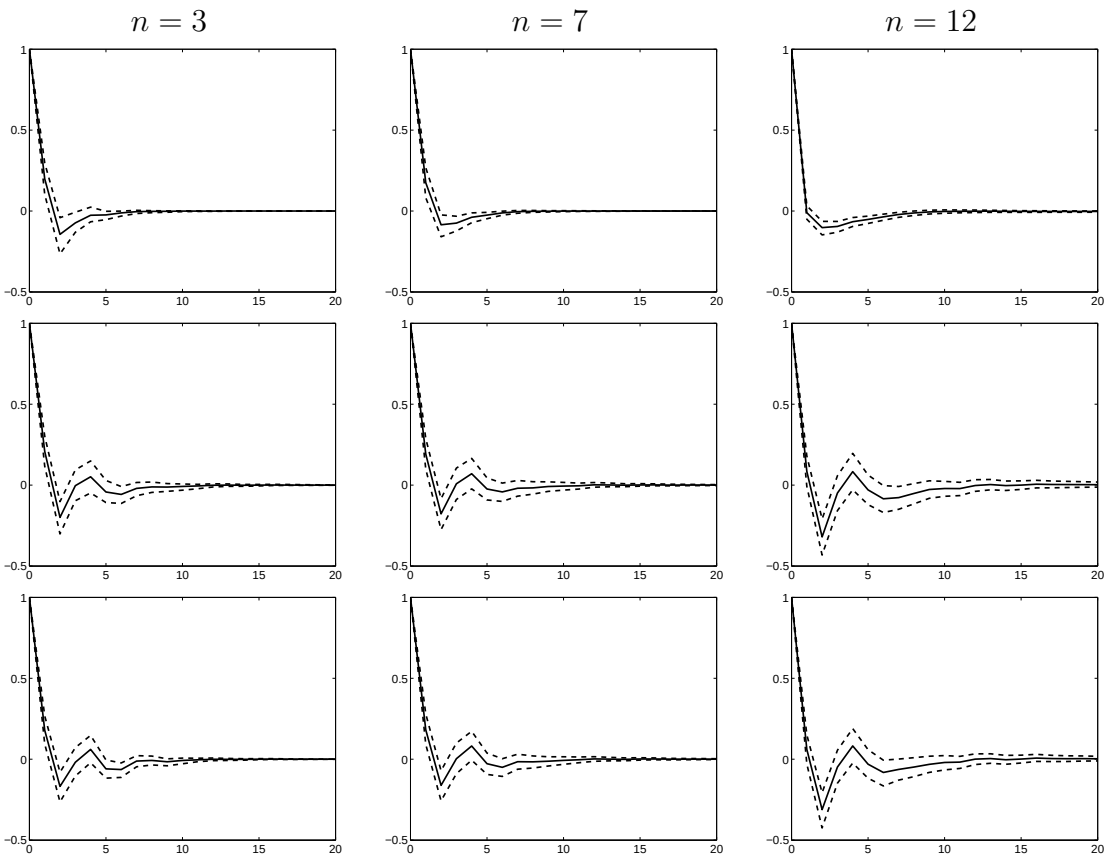

Figure 4: Comparison of impulse responses of the interest rate to own shock. The first, second and third rows contain results for the VARMA$_E(\boldsymbol{\kappa})$, VARMA$(4, 4)$ and VAR(4), respectively. The dotted lines depict the $(10\%, 90\%)$ HPD intervals.

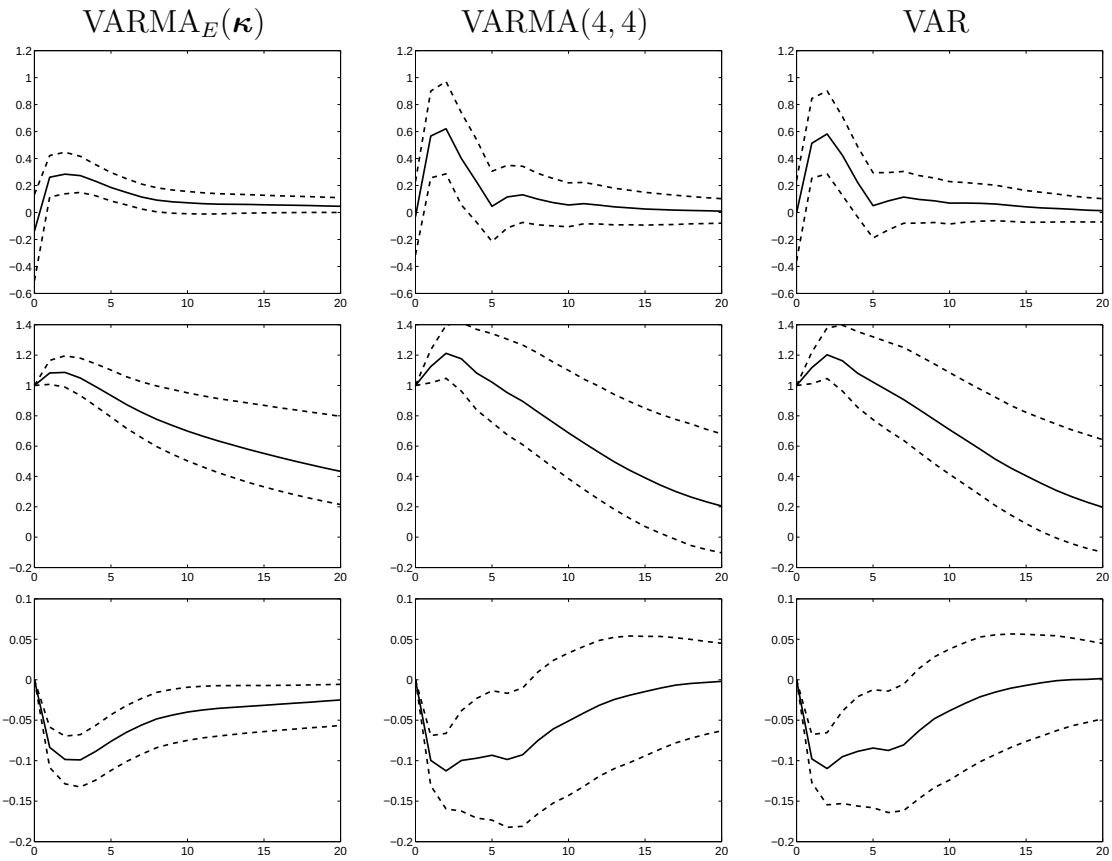

Figure 5: Comparison of impulse responses of the housing start and interest rate to shocks. The first row contains responses of the interest rate to a shock in the housing start; the second row contains responses of the housing start to its own shock; the third row contains responses of the housing start to a shock in the interest rate. The dotted lines depict the $(10\%, 90\%)$ HPD intervals.

sponse of housing starts to an increase in the interest rate. That is, after 20 quarters the model predicts a negative impact with a high degree of certainty (e.g. the HPD interval is all below zero), but one that is very small in magnitude—i.e., approximately -0.025 on average after 20 quarters.

We do not get quite the same picture by looking at the responses generated with the VAR(4). This is mainly due to the larger degree of imprecision induced by the VAR specification. For instance, the impulse response of the interest rate to a shock in housing starts is approximately zero after 20 quarters, with the HPD interval covering both positive and negative regions. Also, the impulse response of housing starts to its own shock is initially large under the VAR, but then falls faster than what the VARMA predicts. At the same time, the VAR generates responses of housing starts to an increase in the interest rate such that the median response vanishes by the end of the 20 quarter horizon. This indicates that an increase in the interest rate has no long term effect on the housing starts, although the HDP intervals are substantially wider than those produced by the VARMA.

All of the approaches discussed so far in this sub-section have included shrinkage using SSVS priors. If we do not include such shrinkage, impulse responses become even more irregular and HPD intervals become even wider. For the sake of brevity, we will not produce conventional impulse responses similar to Figures 2 through 4 for VARMAs and VARs without shrinkage. Suffice it to note here that there is an appreciable deterioration in impulse responses relative to Figures 2 through 4. Instead Figure 6 presents impulse responses relating to the housing variable for $n = 12$. Relative to VARMA$_E(\boldsymbol{\kappa})$ both the VARMA$(4, 4)$ and VAR(4) are producing impulse responses which are much more erratic and with much wider HPD intervals.

In sum, the specification, identification and shrinkage issues investigated in this paper can have an important impact on policy-relevant issues.

5 Conclusions

We began this paper by arguing that there might be some benefits to working with VARMAs instead of VARs. However, VARMAs are little-used due to problems of identification, over-parameterization and computation. In this paper, we have developed a modelling approach, using SSVS priors on both parameters and identification restrictions, which surmounts these problems. Using artificial data and a substantive macroeconomic application, we show that this modelling approach does work well even in VARMAs of high dimension. It is computationally feasible and yields sensible results which have the potential to lead to different policy conclusions than simpler VAR or alternative VARMA approaches.

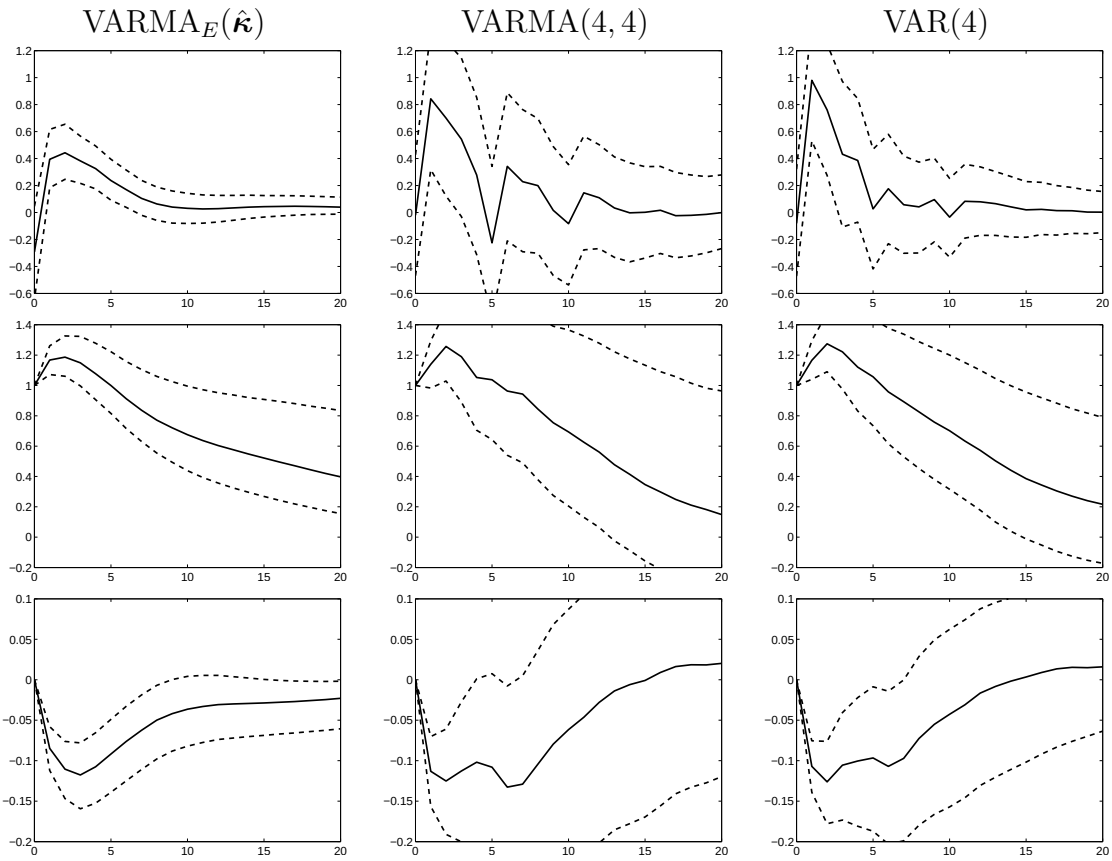

Figure 6: Comparison of impulse responses of the housing start and interest rate to shocks, without using SSVS shrinkage. The first row contains responses of the interest rate to a shock in the housing start; the second row contains responses of the housing start to its own shock; the third row contains responses of the housing start to a shock in the interest rate. The dotted lines depict the $(10\%, 90\%)$ HPD intervals.

References

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

Banbura, M., Giannone, D. and Reichlin, L. (2010). "Large Bayesian vector autoregressions," Journal of Applied Econometrics, 25, 71-92.

Bernanke, B., Boivin, J. and Eliasz, P. (2005). "Measuring monetary policy: a factor augmented autoregressive (FAVAR) approach," Quarterly Journal of Economics, 120, 387-422.

Carriero, A., Clark, T. and Marcellino, M. (2011). "Bayesian VARs: Specification choices and forecast accuracy," Working Paper 1112, Federal Reserve Bank of Cleveland.

Carriero, A., Kapetanios, G. and Marcellino, M. (2009). "Forecasting exchange rates with a large Bayesian VAR," International Journal of Forecasting, 25, 400-417.

Celeux, G., Forbes, F., Robert, C. and Titterington, D. (2006). "Deviance information criteria for missing data models," Bayesian Analysis, 1, 651-674.

Chan, J. (2013). "Moving average stochastic volatility models with application to inflation forecast," Journal of Econometrics, 176, 162-172.

Chan, J. and Eisenstat, E. (2014). "Efficient estimation of Bayesian VARMAs with time-varying coefficients," manuscript.

Chan, J. and Grant, A. (2014). "Fast computation of the deviance information criterion for latent variable models," Computational Statistics and Data Analysis, forthcoming.

Cooley, T. and Dwyer, M. (1998). "Business cycle analysis without much theory. A look at structural VARs," Journal of Econometrics, 83, 57-88.

Dias, G. and Kapetanios, G. (2013). "Forecasting medium and large datasets with Vector Autoregressive Moving Average (VARMA) models," manuscript.

Doan, T., Litterman, R. and Sims, C. (1984). "Forecasting and conditional projection using realistic prior distributions," Econometric Reviews, 3, 1-144.

Fernandez-Villaverde, J., Rubio-Ramirez, J., Sargent, T. and Watson, M. (2007). "A, B, C's (and D's) for understanding VARs," American Economic Review, 97, 1021-1026.

Gefang, D. (2014). "Bayesian doubly adaptive elastic-net lasso for VAR shrinkage," International Journal of Forecasting, 30, 1-11.

George, E., Sun, D. and Ni, S. (2008). "Bayesian stochastic search for VAR model restrictions," Journal of Econometrics, 142, 553-580.

Giannone, D., Lenza, M., Momferatou, D. and Onorante, L. (2010). "Short-term inflation projections: a Bayesian vector autoregressive approach," ECARES working paper 2010-011, Universite Libre de Bruxelles.

Hannan, E. J. (1976). "The identification and parameterization of ARMAX and state space forms, " Econometrica, 44, 713-â€"723.

Koop, G. (2013). "Forecasting with Medium and Large Bayesian VARs," Journal of Applied Econometrics, 28, 177-203.

Koop, G. (2014). "Forecasting with Dimension Switching VARs," International Journal of Forecasting, 30, 280-290.

Korobilis, D. (2013). "VAR forecasting using Bayesian variable selection," Journal of Applied Econometrics, 28, 204-230.

Kuo, L. and Mallick, B. (1997). "Variable selection for regression models," Shankya: Indian Journal of Statistics (Series B), 60, 65–81.

Li, H. and Tsay, R. (1998). "A unified approach to identifying multivariate time series models," 93, 770-782.

Litterman, R. (1986). "Forecasting with Bayesian vector autoregressions – Five years of experience," Journal of Business and Economic Statistics, 4, 25-38.

Lutkepohl, H. (2005). New Introduction to Multiple Time Series Analysis. Springer-Verlag, Berlin.

Lutkephohl, H. and Poskitt, D. (1996). "Specification of echelon form VARMA models," Journal of Business and Economic Statistics, 14, 69-79.

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.

Poskitt, D. (1992). "Identification of echelon canonical forms for vector linear processes using least squares," Annals of Statistics, 20, 195-215.

Primiceri., G. (2005). "Time varying structural vector autoregressions and monetary policy," Review of Economic Studies, 72, 821-852.

Ravishanker, N. and Ray, B. (1997). "Bayesian analysis of vector ARMA models using Gibbs sampling," Journal of Forecasting, 16, 177-194

Sims, C. (1980). "Macroeconomics and reality," Econometrica, 48, 1-48.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002). "Bayesian measures of model complexity and fit," Journal of the Royal Statistical Society Series B, 64, 583-639.

Stock, J. and Watson, M. (2008). "Forecasting in dynamic factor models subject to structural instability," in *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry*, edited by J. Castle and N. Shephard, Oxford: Oxford University Press.

Appendix A: Priors

The empirical work in this paper uses relatively noninformative priors. The hierarchical SSVS priors for the VARMA coefficients are described in sub-section 2.2. Recall that in terms of these, we specify uniform priors on the Kronecker indices $\boldsymbol{\kappa}$. Moreover, for both the hard and soft SSVS priors, we set $\tau^2_{1,ijk} = 1$; for soft SSVS we set $\tau^2_{0,ijk} = 0.01$.

The remaining parameters are assigned the following priors:

$$\begin{aligned}
\Lambda_{i,i} &\sim \mathcal{IG}(\nu_{\lambda,0}, S_{\lambda,0}), \\
\Omega_{i,i} &\sim \mathcal{IG}(\nu_{\omega,0}, S_{\omega,0}), \\
h_{i,0} &\sim \mathcal{N}(h_{0,0}, V_{h,0}), \\
\sigma^2_{h,i} &\sim \mathcal{IG}(\nu_{h,0}, S_{h,0}).
\end{aligned}$$

We set $\nu_{\lambda,0} = 0$, $S_{\lambda,0} = 0.1$, which implies an improper prior on $\Lambda_{i,i}$, and $\nu_{\omega,0} = 5$, $S_{\lambda,0} = 0.4$, $\nu_{h,0} = 5$, $S_{\lambda,0} = 0.4$, $h_{0,0} = 0$, $V_{h,0} = 10$.

Appendix B: MCMC Algorithm

We write the model as

$$\mathbf{y}_t = \mathbf{B}\mathbf{X}_t + \mathbf{\Phi}\mathbf{F}_t + \boldsymbol{\eta}_t, \tag{10}$$

where $\mathbf{B} = (\mathrm{I}_n - \mathbf{B}_0, \mathbf{B}_1, \ldots, \mathbf{B}_p)$, $\mathbf{\Phi} = (\mathbf{\Phi}_0, \ldots, \mathbf{\Phi}_q)$, $\mathbf{X}_t = (\mathbf{y}_t', \ldots, \mathbf{y}_{t-p}')'$ and $\mathbf{F}_t = (\mathbf{f}_t', \ldots, \mathbf{f}_{t-q}')'$. Note that this nests both the expanded and echelon form VARMAs. For notational convenience, define the vector of row degrees $\mathbf{p} = (p_1, \ldots, p_n)'$. The model parameters are sampled using a Gibbs sampler consisting of the following steps:

1. Sample $\left(\boldsymbol{p} \,|\, \boldsymbol{\gamma}^S, \mathbf{f}, \mathbf{\Lambda}\right)$ *marginal* of $\mathbf{B}, \mathbf{\Phi}$ and compute $\boldsymbol{\gamma}^R$ as $\gamma^{B,R}_{ijk} = \gamma^{\Phi,R}_{ijk} = 1$ *iff* $0 < j \leq \rho_i$ or $j = 0, i < k$. This is done with a multi-move sampler that draws $\left(p_i \,|\, \boldsymbol{p}_{-i}, \boldsymbol{\gamma}^S, \mathbf{f}, \mathbf{\Lambda}\right)$ for each $i = 1, \ldots, n$. For the exact echelon algorithm set $\kappa_i = p_i$. To sample p_i, we compute the weights $\mathbb{P}(p_i = l \,|\, \cdot\,)$ using the conditional likelihood $p(\mathbf{y}_i \,|\, \boldsymbol{\gamma}^{R,l}, \boldsymbol{\gamma}^S, \mathbf{f}, \Lambda_{i,i})$ where $\boldsymbol{\gamma}^{R,l}$ are the restrictions implied by $p_1, \ldots, l, \ldots, p_n$.

   To evaluate each conditional likelihood, observe that conditional on $\mathbf{f}$, the model maybe written as n independent regressions. Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_T)'$, $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_T)'$ and set $\mathbf{W} = (\mathbf{X}, \mathbf{F})$. Then,

$$\mathbf{y}_i = \mathbf{W}\boldsymbol{\delta}_i + \boldsymbol{\eta}_i, \qquad \boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \Lambda_{i,i}\mathrm{I}_T), \tag{11}$$

   where $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,T})'$ and $\boldsymbol{\delta}_i$ is the i-th column of $(\mathbf{B}, \mathbf{\Phi})'$.

   Now, a given set of restrictions $\boldsymbol{\gamma}^{R,l}$ will force certain elements in $\boldsymbol{\delta}_i$ to be zero. Define $\boldsymbol{\delta}_i^\star$ to be the vector containing only the *free* elements of $\boldsymbol{\delta}_i$ and $\mathbf{W}_i^\star$ the matrix $\mathbf{W}$ with column $\mathbf{W}_k$ removed for any $\delta_{i,k} = 0$. Clearly, $\mathbf{W}\boldsymbol{\delta} = \mathbf{W}_i^\star\boldsymbol{\delta}_i^\star$ and

$$\left(\boldsymbol{\delta}_i^\star \,|\, \boldsymbol{\gamma}^S\right) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}^\star_{\delta_i,0}),$$

   if "soft" SSVS priors are specified on B_{ijk}, Φ_{ijk}. In this case, $\mathbf{V}^\star_{\delta_i,0}$ is a diagonal matrix with element $V^\star_{\delta_i,0,l,l} = \tau^2_{0,ijk}$ (i.e. the "small" variance) if $\delta^\star_{i,l}$ corresponds to either B_{ijk} with $\gamma^{B,S}_{ijk} = 0$ or to Φ_{ijk} with $\gamma^{\Phi,S}_{ijk} = 0$. Otherwise, $V^\star_{\delta_i,0,l,l} = \tau^2_{1,ijk}$ (i.e. the "large" variance).

   It is straightforward to show in this case that

$$\begin{aligned}(\mathbf{y}_i \,|\, \boldsymbol{\gamma}^{R,l}, \boldsymbol{\gamma}^S, \mathbf{f}, \Lambda_{i,i}) &\sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{V}}_{y_i}),\\ \hat{\mathbf{V}}_{y_i} &= \left(\Lambda_{i,i}^{-1}\mathrm{I}_T - \Lambda_{i,i}^{-2}\mathbf{W}_i^\star\hat{\mathbf{\Delta}}_i^{-1}\mathbf{W}_i^{\star\prime}\right)^{-1},\\ \hat{\mathbf{\Delta}}_i &= \mathbf{V}^{\star-1}_{\delta_i,0} + \Lambda_{i,i}^{-1}\mathbf{W}_i^{\star\prime}\mathbf{W}_i^\star.\end{aligned} \tag{12}$$

   The quadratic term $\mathbf{y}_i'\hat{\mathbf{V}}^{-1}_{y_i}\mathbf{y}_i$ is easy to evaluate (i.e. without the need to separately compute the inverse of $\hat{\mathbf{V}}_{y_i}$), as well as the determinant $|\hat{\mathbf{V}}_{y_i}| = \Lambda_{i,i}^T|\mathbf{V}^\star_{\delta_i,0}||\hat{\mathbf{\Delta}}_i|$. Therefore, computing the likelihood ratio in (12) entails little computation difficulty.

   To evaluate the likelihood ratio under the "hard" SSVS prior, define $\mathbf{W}_i^\circ$ as the matrix $\mathbf{W}_i^\star$ with the l-th column removed for every $\delta^\star_{i,l}$ that corresponds to either

B_{ijk} with $\gamma^{B,S}_{ijk} = 0$ or to Φ_{ijk} with $\gamma^{\Phi,S}_{ijk} = 0$. Also, let $\mathbf{V}^{\circ}_{\delta_i,0}$ be the prior covariance for the unrestricted elements in $\boldsymbol{\delta}_i$. The conditional likelihood is now

$$\begin{aligned}(\mathbf{y}_i \,|\, \boldsymbol{\gamma}^{R,l}, \boldsymbol{\gamma}^{S}, \mathbf{f}, \Lambda_{i,i}) &\sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{V}}_{y_i}), \\ \hat{\mathbf{V}}_{y_i} &= \left(\Lambda^{-1}_{i,i}\mathrm{I}_T - \Lambda^{-2}_{i,i}\mathbf{W}^{\circ}_i \hat{\boldsymbol{\Delta}}_i^{-1} \mathbf{W}^{\circ\prime}_i\right)^{-1}, \\ \hat{\boldsymbol{\Delta}}_i &= \mathbf{V}^{\circ -1}_{\delta_i,0} + \Lambda^{-1}_{i,i}\mathbf{W}^{\circ\prime}_i \mathbf{W}^{\circ}_i, \end{aligned} \tag{13}$$

and computation is similarly straightforward.

Now, to enforce the echelon form in an exact manner, we need to take into account the prior in (7). Practically, this means computing the indicators $\boldsymbol{\gamma}^{E,l} = \mathcal{E}(p_1, \ldots, l, \ldots, p_n)$ and setting the weights:

$$\mathbb{P}(p_i = l \,|\, \cdot\,) \propto \begin{cases} 0 \text{ if } \gamma^{B,E,l}_{ijk} = 0, \gamma^{B,R,l}_{ijk} \neq 0, \gamma^{B,S}_{ijk} \neq 0 \text{ for any } j,k \\ p(\mathbf{y}_i \,|\, \boldsymbol{\gamma}^{R,l}, \boldsymbol{\gamma}^{S}, \mathbf{f}, \Lambda_{i,i}) \text{ otherwise} \end{cases}. \tag{14}$$

For the approximate row degree algorithm, however, the above step is skipped and we simply set:

$$\mathbb{P}(p_i = l \,|\, \cdot\,) \propto p(\mathbf{y}_i \,|\, \boldsymbol{\gamma}^{R,l}, \boldsymbol{\gamma}^{S}, \mathbf{f}, \Lambda_{i,i}). \tag{15}$$

Observe that in this case, not only do we circumvent the need to compute echelon form indicators and check them against the row degree and SSVS indicators, but also p_i is conditionally independent of the other row degrees $\mathbf{p}_{-i}$. The approximate algorithm, therefore, is both computationally simpler and more efficient (albeit at the cost of loosing the exact canonical form).

2. Sample $\left(\boldsymbol{\gamma}^{S}_i, \mathbf{B}_i, \boldsymbol{\Phi}_i \,|\, \boldsymbol{\gamma}^{R}, \mathbf{f}, \Lambda_{i,i}, \mathbf{y}_i\right)$ for each $i = 1, \ldots, n$, where $\mathbf{B}_i$ denotes the i-th row of $\mathbf{B}$, $\boldsymbol{\Phi}_i$ the i-th row of $\boldsymbol{\Phi}$, and $\boldsymbol{\gamma}^{S}_i$ is the set of all SSVS indicators pertaining to $\mathbf{B}_i, \boldsymbol{\Phi}_i$. Under the "hard" SSVS prior, this is done by first sampling $\left(\boldsymbol{\gamma}^{S}_i, \,|\, \boldsymbol{\gamma}^{R}, \mathbf{f}, \Lambda_{i,i}\right)$ marginal of $\mathbf{B}_i, \boldsymbol{\Phi}_i$ using (11). In particular, we sample each $\gamma^{\cdot,S}_{ijk}$ for every j,k conditional on $\{\gamma^{\cdot,S}_{i,lm}\}_{l\neq j, m\neq k}$, using the approach outlined in Step 1 to compute the likelihood ratio

$$\varrho^{\cdot,S}_{ijk} = \frac{p(\mathbf{y}_i \,|\, \boldsymbol{\gamma}^{R}, \gamma^{\cdot,S}_{ijk} = 1, \{\gamma^{\cdot,S}_{ilm}\}_{l\neq j, m\neq k}, \mathbf{f}, \Lambda_{i,i})}{p(\mathbf{y}_i \,|\, \boldsymbol{\gamma}^{R}, \gamma^{\cdot,S}_{ijk} = 0, \{\gamma^{\cdot,S}_{ilm}\}_{l\neq j, m\neq k}, \mathbf{f}, \Lambda_{i,i})}. \tag{16}$$

Given our priors, this implies

$$\mathbb{P}\left(\gamma^{\Phi,S}_{ijk} = 1 \,|\, \boldsymbol{\gamma}^{R}, \{\gamma^{\Phi,S}_{ilm}\}_{l\neq j, m\neq k}, \mathbf{f}, \Lambda_{i,i}, \mathbf{y}_i\right) = \varrho^{\Phi,S}_{ijk} / \left(1 + \varrho^{\Phi,S}_{ijk}\right), \tag{17}$$

for both the exact echelon form algorithm and the approximate row degrees algorithm. For the indicators on $\mathbf{B}_i$, however, imposing the echelon form once again requires that the relationship between $\boldsymbol{\gamma}^{B,S}_i$ and $\boldsymbol{\gamma}^{B,R}$ established in (7) be respected. In consequence, the correct conditional distribution is given by

$$\begin{aligned} &\mathbb{P}\left(\gamma^{B,S}_{ijk} = 1 \,|\, \boldsymbol{\gamma}^{R}, \{\gamma^{B,S}_{ilm}\}_{l\neq j, m\neq k}, \mathbf{f}, \Lambda_{i,i}, \mathbf{y}_i\right) \\ &= \begin{cases} 0 & \text{if } \gamma^{B,E}_{ijk} = 0, \gamma^{B,R}_{ijk} \neq 0 \\ \varrho^{B,S}_{ijk} / \left(1 + \varrho^{B,S}_{ijk}\right) & \text{otherwise} \end{cases}, \end{aligned} \tag{18}$$

where $\gamma_{ijk}^{B,E}$ is computed from $\mathcal{E}(p_1,\ldots,p_n)$ using previous draws of $p_1,\ldots,p_n$. For the approximate row degrees algorithm, however, we simply draw from

$$\mathbb{P}\left(\gamma_{ijk}^{B,S}=1\,|\,\boldsymbol{\gamma}^{R},\{\gamma_{ilm}^{B,S}\}_{l\neq j,m\neq k},\mathbf{f},\Lambda_{i,i},\mathbf{y}_i\right)=\varrho_{ijk}^{B,S}/\left(1+\varrho_{ijk}^{B,S}\right). \tag{19}$$

For sake of efficient computation, we note that whenever $\gamma_{ijk}^{\cdot,R}=0$, we always obtain $\varrho_{ijk}^{\cdot,S}=1$, and therefore, the conditional likelihoods need not be computed.

When using the "soft" SSVS prior, the indicators are sampled conditional on $\mathbf{B}_i,\boldsymbol{\Phi}_i$. For $\gamma_{ijk}^{B,S}$, if the echelon form is imposed, then $\mathbb{P}\left(\gamma_{ijk}^{B,S}=1\,|\,B_{ijk}\right)=0$ when $\boldsymbol{\gamma}_{ijk}^{B,E}=0$ and $\boldsymbol{\gamma}_{ijk}^{B,R}\neq 0$; otherwise

$$\mathbb{P}\left(\gamma_{ijk}^{B,S}=1\,|\,B_{ijk}\right)=\frac{\frac{1}{\tau_{1,ijk}}\exp\left(-\frac{B_{ijk}^2}{2\tau_{1,ijk}^2}\right)}{\frac{1}{\tau_{1,ijk}}\exp\left(-\frac{B_{ijk}^2}{2\tau_{1,ijk}^2}\right)+\frac{1}{\tau_{0,ijk}}\exp\left(-\frac{B_{ijk}^2}{2\tau_{0,ijk}^2}\right)}. \tag{20}$$

If the row degree algorithm is used, we sample $\gamma_{ijk}^{B,S}$ using only (20).

For $\gamma_{ijk}^{\Phi,S}$, the success probabilities are

$$\mathbb{P}\left(\gamma_{ijk}^{\Phi,S}=1\,|\,\Phi_{ijk},\boldsymbol{\gamma}_{ijk}^{\Phi,R}\neq 0\right)=\frac{\frac{1}{\tau_{1,ijk}}\exp\left(-\frac{\Phi_{ijk}^2}{2\tau_{1,ijk}^2}\right)}{\frac{1}{\tau_{1,ijk}}\exp\left(-\frac{\Phi_{ijk}^2}{2\tau_{1,ijk}^2}\right)+\frac{1}{\tau_{0,ijk}}\exp\left(-\frac{\Phi_{ijk}^2}{2\tau_{0,ijk}^2}\right)}.$$

Once again, $\gamma_{ijk}^{\cdot,S}$ is drawn from $\mathbb{P}(\gamma_{ijk}^{\cdot,S}=1\,|\,\gamma_{ijk}^{\cdot,R}=0)=0.5$ whenever the corresponding coefficient is excluded by the row degrees.

Given a draw of $\boldsymbol{\gamma}_i^S$, the coefficients $\mathbf{B}_i,\boldsymbol{\Phi}_i$ are sampled jointly in standard way for both of the SSVS specifications. In particular, letting once again $\mathbf{W}_i^\star=\begin{pmatrix}\mathbf{X}^\star & \mathbf{F}^\star\end{pmatrix}$ be the reduced regressors and factors matrix corresponding to the unrestricted coefficients $\boldsymbol{\delta}_i^\star$ in $\boldsymbol{\delta}_i$, textbook regression analysis dictates

$$\begin{aligned}\left(\boldsymbol{\delta}_i^\star\,|\,\boldsymbol{\gamma}_i^S,\boldsymbol{\gamma}_i^R,\mathbf{f},\Lambda_{i,i},\mathbf{y}_i\right)&\sim\mathcal{N}(\hat{\boldsymbol{\delta}}_i,\hat{\boldsymbol{\Delta}}_i),\\ \hat{\boldsymbol{\delta}}_i&=\hat{\boldsymbol{\Delta}}_i\left(\Lambda_{i,i}^{-1}\mathbf{W}_i^{\star\prime}(\mathbf{y}_i-\mathbf{f}_i)\right),\\ \hat{\boldsymbol{\Delta}}_i&=\left(\mathbf{V}_{\delta_i,0}^\star+\Lambda_{i,i}^{-1}\mathbf{W}_i^{\star\prime}\mathbf{W}_i^\star\right)^{-1},\end{aligned} \tag{21}$$

where $\mathbf{f}_i=(f_{i,1},\ldots,f_{i,T})'$. The remaining elements in $\boldsymbol{\delta}_i$ (and therefore $\mathbf{B}_i,\boldsymbol{\Phi}_i$) are set to zero.

3. Sample

$$\left(\Lambda_{i,i}\,|\,\mathbf{B}_i,\boldsymbol{\Phi}_i,\boldsymbol{\gamma}_i^R,\boldsymbol{\gamma}_i^S,\mathbf{f},\mathbf{y}_i\right)\sim\mathcal{IG}\left(\nu_{\lambda,0}+\frac{T}{2},S_{\lambda,0}+\frac{1}{2}\sum_{t=1}^{T}\left(y_{i,t}-\mathbf{B}_i\mathbf{X}_t-\boldsymbol{\Phi}_i\mathbf{F}_t\right)^2\right)$$

for each $i=1,\ldots,n$.

4. Sample $(\Omega_{i,i} \,|\, \mathbf{f}_i)$ or $\left(h_{i,0}, \ldots, h_{i,T}, \sigma^2_{h,i} \,|\, \mathbf{f}_i\right)$—depending on whether stochastic volatility is specified—for each $i = 1, \ldots, n$. In either case, standard methods are used.

5. Sample $\left(\mathbf{f} \,|\, \mathbf{B}, \boldsymbol{\Phi}, \tilde{\boldsymbol{\Omega}}, \boldsymbol{\Lambda}, \boldsymbol{\gamma}^R, \boldsymbol{\gamma}^S, \mathbf{y}\right)$, where $\tilde{\boldsymbol{\Omega}} = \mathrm{I}_T \otimes \boldsymbol{\Omega}$ for the constant variance case and

$$\tilde{\boldsymbol{\Omega}} = \mathrm{diag}(\exp h_{1,1}, \ldots, \exp h_{n,1}, \ldots, \exp h_{1,T}, \ldots, \exp h_{n,T})$$

for stochastic volatility. An efficient sampler for this purpose is constructed by first rewriting the working model (10) in stacked form as

$$\mathbf{y}^* = \boldsymbol{\Psi}\mathbf{f} + \boldsymbol{\eta}, \tag{22}$$

where $\mathbf{y}^* = ((\mathbf{y}_1 - \mathbf{B}\mathbf{X}_1)', \ldots, (\mathbf{y}_T - \mathbf{B}\mathbf{X}_T)')'$ and $\boldsymbol{\Psi}$ is a $Tn \times Tn$ lower triangular matrix with $\boldsymbol{\Phi}_0$ on the main diagonal block, $\boldsymbol{\Phi}_1$ on first lower diagonal block, $\boldsymbol{\Phi}_2$ on second lower diagonal block, and so forth. For example, for $q = 2$, we have

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Phi}_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{\Phi}_2 & \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_2 & \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_0 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Phi}_2 & \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_0 \\ & & & & & \end{pmatrix}.$$

It is important to note that in general $\boldsymbol{\Psi}$ is a sparse $Tn \times Tn$ matrix that contains at most

$$n^2 \left((q+1)T - \frac{q(q+1)}{2} \right) < n^2(q+1)T$$

non-zero elements, which grows *linearly* in T and is substantially less than the total $(Tn)^2$ elements for typical applications where $T \gg q$.

Now, the vector of factors is sampled jointly as

$$\begin{aligned} \left(\mathbf{f} \,|\, \mathbf{B}, \boldsymbol{\Phi}, \boldsymbol{\Omega}_{(t)}, \boldsymbol{\Lambda}, \boldsymbol{\gamma}^R, \boldsymbol{\gamma}^S, \mathbf{y}\right) &\sim \mathcal{N}(\hat{\mathbf{f}}, \hat{\mathbf{V}}_f), \\ \hat{\mathbf{f}} &= \hat{\mathbf{V}}_f \left(\boldsymbol{\Psi}' \left(\mathrm{I}_T \otimes \boldsymbol{\Lambda}^{-1} \right) \mathbf{y}^* \right), \\ \hat{\mathbf{V}}_f &= \left(\tilde{\boldsymbol{\Omega}}^{-1} + \boldsymbol{\Psi}' \left(\mathrm{I}_T \otimes \boldsymbol{\Lambda}^{-1} \right) \boldsymbol{\Psi} \right)^{-1}, \end{aligned} \tag{23}$$

which is once again efficiently implemented using sparse matrix routines.

Appendix C: Deviance Information Criterion

The Deviance Information Criterion (DIC) was introduced in Spiegelhalter, Best, Carlin, and van der Linde (2002). For latent variable models there are numerous definitions (Celeux, Forbes, Robert, and Titterington, 2006) depending on the exact notion of the likelihood. Given a likelihood function $f(\mathbf{y} \,|\, \boldsymbol{\theta})$, the DIC is defined as:

$$\text{DIC} = \overline{D(\boldsymbol{\theta})} + p_D,$$

where

$$\overline{D(\boldsymbol{\theta})} = -2\mathbb{E}_{\boldsymbol{\theta}}[\log f(\mathbf{y} \,|\, \boldsymbol{\theta}) \,|\, \mathbf{y}]$$

is the posterior mean deviance and p_D is the effective number of parameters. That is, the DIC is the sum of the posterior mean deviance, which can be used as a Bayesian measure of model fit or adequacy, and the effective number of parameters that measures model complexity. The effective number of parameters is in turn defined as

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\widetilde{\boldsymbol{\theta}}),$$

where $D(\boldsymbol{\theta}) = -2\log f(\mathbf{y} \,|\, \boldsymbol{\theta})$ and $\widetilde{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$, which is typically taken as the posterior mean.

Our VARMA model has a few equivalent latent variable representations. Hence, in principle we can use any of the representations and compute the DIC based on the conditional likelihood (i.e., the likelihood given the latent variables). However, as pointed out in Chan and Grant (2014), conditional DICs tend to be numerically unstable. Instead, we use the likelihood implied by the system

$$\mathbf{y}_t = \sum_{j=1}^{p} \mathbf{A}_j \mathbf{y}_{t-j} + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{\epsilon}_{t-j} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{24}$$

where all the parameters are identified and can be recovered from the main sampling algorithm.

To derive this density, we stack (24) over t and obtain:

$$\mathbf{y} = \mathbf{a} + \boldsymbol{\Theta}\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', \dots, \boldsymbol{\epsilon}_T')' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_T \otimes \boldsymbol{\Sigma})$, $\mathbf{a} = ((\sum_{j=1}^{p} \mathbf{A}_j \mathbf{y}_{1-j})', \dots, (\sum_{j=1}^{p} \mathbf{A}_j \mathbf{y}_{T-j})')'$ and $\boldsymbol{\Theta}$ is a $Tn \times Tn$ lower triangular matrix with the identity matrix $\mathbf{I}_n$ on the main diagonal block, $\boldsymbol{\Theta}_1$ on first lower diagonal block, $\boldsymbol{\Theta}_2$ on second lower diagonal block, and so forth. Hence, we have

$$(\mathbf{y} \,|\, \mathbf{A}_1, \dots, \mathbf{A}_p, \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}) \sim \mathcal{N}(\mathbf{a}, \boldsymbol{\Theta}(\mathbf{I}_T \otimes \boldsymbol{\Sigma})\boldsymbol{\Theta}').$$

Since the covariance matrix $\boldsymbol{\Theta}(\mathbf{I}_T \otimes \boldsymbol{\Sigma})\boldsymbol{\Theta}'$ is a band matrix, this Normal density can be evaluated quickly using the band matrix algorithms discussed in Chan and Grant (2014).

Appendix D: Data Appendix

All variables were downloaded from St. Louis' FRED database and cover the quarters 1959:Q1 to 2013:Q4. The following table lists the variables, describes how they were transformed and whether they are slow- or fast-moving variables. The transformation codes are: 1 - no transformation (levels); 2 - first difference, 3 - second difference; 4 - logarithm; 5 - first difference of logarithm; 6 - second difference of logarithm.

	Trans.	Slow /	included in model		
Variable	Code	Fast	$n = 3$	$n = 7$	$n = 12$
Real Gross Domestic Product	5	S	X	X	X
Consumer Price Index: All Items	6	S	X	X	X
Real Personal Consumption Exp.	5	S			X
Housing Starts: Total	4	S			X
Average Hourly Earnings: Manuf.	6	S		X	X
Real Gross Private Domestic Invest.	5	S			X
All Employees: Total nonfarm	5	S			X
ISM Manuf.: PMI Composite Index	1	S			X
Effective Federal Funds Rate	2		X	X	X
S&P 500 Stock Price Index	5	F			X
M2 Money Stock	6	F		X	X
Spot Oil Price: West Texas Interm.	5	F		X	X