Efficient Estimation of Bayesian VARMAs with Time-Varying Coefficients

Joshua C.C. Chan*
Economics Discipline Group,
University of Technology Sydney

Eric Eisenstat†
School of Economics,
The University of Queensland

March 2017

Abstract

Empirical work in macroeconometrics has been mostly restricted to using VARs, even though there are strong theoretical reasons to consider general VARMAs. A number of articles in the last two decades have conjectured that this is because estimation of VARMAs is perceived to be challenging and proposed various ways to simplify it. Nevertheless, VARMAs continue to be largely dominated by VARs, particularly in terms of developing useful extensions. We address these computational challenges with a Bayesian approach. Specifically, we develop a Gibbs sampler for the basic VARMA, and demonstrate how it can be extended to models with time-varying VMA coefficients and stochastic volatility. We illustrate the methodology through a macroeconomic forecasting exercise. We show that in a class of models with stochastic volatility, VARMAs produce better density forecasts than VARs, particularly for short forecast horizons.

Keywords: state space, stochastic volatility, factor model, macroeconomic forecasting, density forecast

JEL codes: C11, C32, C53

*Joshua Chan would also like to acknowledge financial support by the Australian Research Council via a Discovery Early Career Researcher Award (DE150100795).

†Corresponding author: School of Economics, The University of Queensland, Level 6, Colin Clark Building (39), St. Lucia, Brisbane Qld 4072 (Australia), e.eisenstat@uq.edu.au.

1 Introduction

Vector autoregressive moving average (VARMA) models have long been considered an appropriate framework for modeling covariance stationary time series. As is well known, by the Wold decomposition theorem, any covariance stationary time series has an infinite moving average representation. Whenever this is characterized by a rational transfer function, the multivariate series can be exactly represented by a finite-order VARMA model. When the transfer function is irrational, the VARMA specification can be used to provide an arbitrarily close approximation (Lütkepohl and Poskitt, 1996).

However, in most empirical work, only purely autoregressive models are considered. In fact, since the seminal work of Sims (1980), VARs have become the most prominent approach in empirical macroeconometrics. This is in spite of long-standing criticisms of VARs, especially with short lag orders, as being theoretically deficient for macroeconomic applications. There are two main theoretical drawbacks of VARs in this context: first, linearized DSGE models typically result in VARMAs, not VARs (e.g., Cooley and Dwyer, 1998; Yang, 2005; Fernández-Villaverde, Rubio-Ramírez, Sargent, and Watson, 2007; Leeper, Walker, and Yang, 2008). Second, even if a particular set of variables can be adequately described by a VAR, any linear combination, temporal aggregation, or subsets of these variables will follow a VARMA process.

Over the past two decades, a number of authors (e.g., Lütkepohl and Poskitt, 1996; Lütkepohl and Claessen, 1997; Athanasopoulos and Vahid, 2008; Athanasopoulos, Poskitt, and Vahid, 2012; Dufour and Stevanović, 2013; Dufour and Pelletier, 2014; Kascha and Trenkler, 2014; Poskitt, 2016) have pointed out this unfortunate phenomenon and various approaches have been proposed aimed at making VARMAs accessible to applied macroeconomists. Nevertheless, VARs continue to dominate in this field. One possible reason for this is that many flexible extensions of the basic VAR have been developed. For example, VARs are now routinely augmented with time-varying coefficients (Canova, 1993; Koop and Korobilis, 2013), Markov switching or regime switching processes (Paap and van Dijk, 2003; Koop and Potter, 2007), and stochastic volatility (Cogley and Sargent, 2005; Primiceri, 2005). Recently, specifications targeting a large number of variables such as factor augmented VARs (Bernanke, Boivin, and Eliasz, 2005; Korobilis, 2012), among many others (see, e.g., Koop and Korobilis, 2010) have been introduced.

While these extensions have made VARs extremely flexible, papers such as Banbura, Giannone, and Reichlin (2010); Koop (2011); Korobilis (2013); Eisenstat, Chan, and Strachan (2016) have focused on achieving parsimony and controlling over-parameterization. This balance between flexibility and parsimony have contributed to the success of VARs in forecasting macroeconomic time series. Seemingly, these developments surrounding VARs have largely overshadowed the advantages inherent to VARMA specifications. Indeed, VARMAs remain largely underdeveloped in this sense, and this is a central concern of the present paper.

The lack of development of VARMAs is unfortunate because well-specified VARMAs are

naturally parsimonious, and recent evidence suggests that VARMAs do forecast macroeconomic variables better than basic VARs (Athanasopoulos and Vahid, 2008; Athanasopoulos, Poskitt, and Vahid, 2012; Kascha and Trenkler, 2014).[1] Therefore, similar extensions such as time-varying coefficients and stochastic volatility in a VARMA specification may offer even further forecasting gains. Moreover, for structural analysis such as estimation of impulse response functions, VARs remain fundamentally deficient. Cooley and Dwyer (1998), and more recently Poskitt and Yao (2016), argue that typical applications of structural VARs in macroeconomics use lag lengths that are much too short to adequately approximate the underlying, theoretically founded VARMA processes. In certain cases—such as in DSGEs with fiscal foresight—the VARMA process arising in equilibrium entails a *non-fundamental* moving average part, and therefore, a VAR approximation does not exist at all (Yang, 2005; Leeper, Walker, and Yang, 2008).

The identification and estimation of VARMAs is, however, far more involved than with VARs. Even for a pure VMA process, the likelihood is highly nonlinear in the parameters, and in Gaussian models identification further requires imposing constraints on the roots of the determinant of the VMA polynomial, which corresponds to a system of nonlinear constraints on VMA parameters. In consequence, estimation using maximum likelihood or Bayesian methods is difficult, and most practical applications rely on approximate methods rather than exact inference from the likelihood (see Kascha, 2012, for a review). Combining VMA with VAR terms gives rise to further problems in terms of specification and identification (see Lütkepohl, 2005, for a textbook treatment), thereby complicating matters even more.

We propose a Bayesian approach that draws on a few recent developments in the state space literature. First, we make use of a convenient state space representation of the VARMA introduced in Metaxoglou and Smith (2007). Specifically, by using the fact that a VMA plus white noise remains a VMA (Peiris, 1988), the authors write a VARMA as a latent factor model, with the unusual feature that lagged factors also enter the current measurement equation. This linear state space form is an *equivalent*, but over-parameterized representation of the original VARMA. To estimate the model in this form, Metaxoglou and Smith (2007) set *ex ante* certain parameters to pre-determined values and estimate the remaining parameters using the EM algorithm based on the Kalman filter.

Our point of departure is to develop an efficient Gibbs sampler for this state space representation of the VARMA. First, we show that the pre-determined parameter restrictions in this case are neither desirable nor necessary—in a Bayesian setting, we work directly with the "unidentified" model and recover the identified VARMA parameters *ex post*. We emphasize from the start and demonstrate below that doing so *does not* require restrictive priors on the coefficients. Another advantage of this approach is that restrictions on roots can be imposed in the post-processing of draws, rather than directly in the sampling scheme. To further accelerate computation, instead of the conventional forward-filtering

[1] Chan (2013) arrives at a similar conclusion in forecasting inflation with univariate MA models with stochastic volatility.

and backward-smoothing algorithms based on the Kalman filter, we make use of the more efficient precision sampler of Chan and Jeliazkov (2009) to simulate the latent factors.

The significance of our contribution lies in the realization that once the basic VARMA can be efficiently estimated via the Gibbs sampler, a wide variety of generalizations, analogous to those mentioned earlier extending the basic VAR, can also be fitted easily using the machinery of Markov chain Monte Carlo (MCMC) techniques. We will focus on a particular generalization of the VARMA: allowing for time-varying VMA coefficients and stochastic volatility. For estimation using Bayesian methods, we focus on Gaussian disturbances, although our algorithms can be readily adopted for other popular distributions, such as Student's t.

Within this scope, we do not address the important issue of specifying a VARMA in canonical form, except to point out that the methods developed below can be readily used to estimate a VARMA in echelon form, conditional on knowledge of the Kronecker indices. An in-depth investigation of Bayesian approaches to specifying and estimating an echelon form VARMA is undertaken in our related work in Chan, Eisenstat, and Koop (2016). We show in the latter that building on the foundation laid out in the present paper, the same extensions may be incorporated in straightforward fashion in the fully canonical echelon form specification (where Kronecker indices are estimated jointly with the model parameters) as well.

In the present work, our main aim is to assess the forecasting potential of VARMAs with time-varying coefficients and stochastic volatility. Specifically, we investigate whether adding moving average components to VARs with stochastic volatility (i.e., a modern, widespread forecasting tool) improves forecasting performance. The sampling algorithm we develop is expressly suitable for this purpose, and we do find that VARMAs with time-varying coefficients and stochastic volatility generate better density forecasts than their VAR counterparts, particularly for inflation over short horizons.

To our knowledge, few attempts have been made to apply Bayesian methods in specifying and estimating VARMA models. Two noteworthy exceptions are Ravishanker and Ray (1997), who consider a hybrid Metropolis-Hastings algorithm for a basic VARMA, and Li and Tsay (1998), who use stochastic search variable selection (SSVS) priors (e.g., George and McCulloch, 1993) to jointly sample the Kronecker indices and coefficients of a VARMA in echelon form. The latter sampler is based on the observation that each equation within a VARMA is a univariate ARMAX, conditional on the other variables in the system. Both of these approaches, however, are computationally intensive and do not provide a convenient framework for incorporating the type of extensions we develop here.

The rest of this article is organized as follows. Section 2 first introduces a state space representation of a VARMA(p, q)—which we term the *expanded VARMA form*—that facilitates efficient estimation, followed by a detailed discussion of the correspondence between this representation and the original VARMA. In Section 3 we consider a general VARMA framework with time-varying coefficients and stochastic volatility. We then

develop a Gibbs sampler for this general model. In Section 4, the methodology is illustrated using a recursive forecasting exercise involving inflation and GDP growth. Lastly, we discuss some future research directions in Section 5.

2 The Expanded VMA Form

In this section we discuss the identification issues in the expanded VMA form and how one can recover the original VMA parameters from those in the expanded VMA form. The theory developed in sub-sections 2.1 and 2.2 assumes the disturbances follow a weak white noise process. When we discuss recovering the original VMA parameters in sub-section 2.3, as well as estimation in the next section, we will assume a stronger condition—the disturbances follow Gaussian distributions.

2.1 The Basic Setup

To build up the general framework, we start with the pure VMA(q) specification:

$$\mathbf{u}_t = \mathbf{\Theta}(L)\boldsymbol{\varepsilon}_t \equiv \mathbf{\Theta}_0\boldsymbol{\varepsilon}_t + \mathbf{\Theta}_1\boldsymbol{\varepsilon}_{t-1} + \cdots + \mathbf{\Theta}_q\boldsymbol{\varepsilon}_{t-q}, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{WN}(0, \mathbf{\Sigma}), \tag{1}$$

where $\mathbf{u}_t$ is $n \times 1$ and all other matrices conform accordingly. $\mathcal{WN}(0, \mathbf{\Sigma})$ denotes a (weak) white noise process with covariance $\mathbf{\Sigma}$, i.e., $\{\boldsymbol{\varepsilon}_t\}$ satisfies $\mathrm{E}(\boldsymbol{\varepsilon}_t) = \mathbf{0}$, $\mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \mathbf{\Sigma}$, and $\mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_{t-s}') = \mathbf{0}$ for all $s \geq 1$. $\mathbf{\Theta}_0$ is assumed to be non-singular, but not necessarily equal to $\mathbf{I}_n$ (as would be relevant in an *echelon form* VARMA specification, for example). The autocovariances generated by this VMA(q) process are given by

$$\mathbf{\Gamma}_j \equiv \mathrm{E}(\mathbf{u}_t\mathbf{u}_{t-j}') = \sum_{l=j}^{q} \mathbf{\Theta}_l\mathbf{\Sigma}\mathbf{\Theta}_{l-j}', \qquad j = 0, \ldots, q.$$

In what follows, it will also be useful to consider a VMA(1) representation of the general VMA(q), defined as

$$\underbrace{\begin{pmatrix} \mathbf{u}_t \\ \mathbf{u}_{t-1} \\ \vdots \\ \mathbf{u}_{t-q+1} \end{pmatrix}}_{\widetilde{\mathbf{u}}_\tau} = \underbrace{\begin{pmatrix} \mathbf{\Theta}_0 & \mathbf{\Theta}_1 & \cdots & \mathbf{\Theta}_{q-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \mathbf{\Theta}_1 \\ & & & \mathbf{\Theta}_0 \end{pmatrix}}_{\widetilde{\mathbf{\Theta}}_0} \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \boldsymbol{\varepsilon}_{t-1} \\ \vdots \\ \boldsymbol{\varepsilon}_{t-q+1} \end{pmatrix}}_{\widetilde{\boldsymbol{\varepsilon}}_\tau} + \underbrace{\begin{pmatrix} \mathbf{\Theta}_q & & & \\ \mathbf{\Theta}_{q-1} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \mathbf{\Theta}_1 & \cdots & \mathbf{\Theta}_{q-1} & \mathbf{\Theta}_q \end{pmatrix}}_{\widetilde{\mathbf{\Theta}}_1} \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_{t-q} \\ \boldsymbol{\varepsilon}_{t-q-1} \\ \vdots \\ \boldsymbol{\varepsilon}_{t-2q+1} \end{pmatrix}}_{\widetilde{\boldsymbol{\varepsilon}}_{\tau-1}}, \tag{2}$$

with $\widetilde{\boldsymbol{\varepsilon}}_\tau \sim \mathcal{WN}(0, \widetilde{\boldsymbol{\Sigma}})$ and $\widetilde{\boldsymbol{\Sigma}} = \mathbf{I}_q \otimes \boldsymbol{\Sigma}$. In this form, the corresponding autovariances may be denoted by

$$\widetilde{\boldsymbol{\Gamma}}_0 = \begin{pmatrix} \boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_1 & \cdots & \boldsymbol{\Gamma}_{q-1} \\ \boldsymbol{\Gamma}_1' & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{\Gamma}_1 \\ \boldsymbol{\Gamma}_{q-1}' & \cdots & \boldsymbol{\Gamma}_1' & \boldsymbol{\Gamma}_0 \end{pmatrix} \quad \text{and} \quad \widetilde{\boldsymbol{\Gamma}}_1 = \begin{pmatrix} \boldsymbol{\Gamma}_q & & & \\ \boldsymbol{\Gamma}_{q-1} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \boldsymbol{\Gamma}_1 & \cdots & \boldsymbol{\Gamma}_{q-1} & \boldsymbol{\Gamma}_q \end{pmatrix}. \tag{3}$$

Writing the VMA(q) this way allows us to work directly with the simplest case, $q = 1$, and generalize the developed concepts and methods to any q through (2)-(3).

We assume throughout that the VMA(q) process is *invertible*, i.e., the characteristic equation

$$\det \boldsymbol{\Theta}(z) \equiv \det(\boldsymbol{\Theta}_0 + \boldsymbol{\Theta}_1 z + \cdots + \boldsymbol{\Theta}_q z^q)$$

has no roots exactly on the unit circle (equivalently, no eigenvalues of $\widetilde{\boldsymbol{\Theta}}_1 \widetilde{\boldsymbol{\Theta}}_0^{-1}$ in (2) are exactly one in modulus). We shall further refer to a VMA process with all roots of $\det \boldsymbol{\Theta}(z)$ strictly outside the unit circle as *fundamental*, and a VMA process with any root strictly inside the unit circle as *non-fundamental*. Note that a *fundamental* VMA process is invertible in the past of $\mathbf{u}_t$ (i.e. $\mathbf{u}_t, \mathbf{u}_{t-1}, \mathbf{u}_{t-2}, \dots$), while a *non-fundamental* process is invertible in the past and future of $\mathbf{u}_t$ (i.e. $\dots, \mathbf{u}_{t+2}, \mathbf{u}_{t+1}, \mathbf{u}_t, \mathbf{u}_{t-1}, \mathbf{u}_{t-2}, \dots$). The methods we develop below are applicable to both fundamental and non-fundamental processes, as long as they are invertible.

Following Metaxoglou and Smith (2007), consider now the decomposition of $\mathbf{u}_t$:

$$\mathbf{u}_t = \boldsymbol{\Phi}_0 \mathbf{f}_t + \cdots + \boldsymbol{\Phi}_q \mathbf{f}_{t-q} + \boldsymbol{\eta}_t, \qquad \begin{pmatrix} \mathbf{f}_t \\ \boldsymbol{\eta}_t \end{pmatrix} \sim \mathcal{WN}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda} \end{pmatrix} \right), \tag{4}$$

where $\mathbf{f}_t$ is $n \times 1$, $\boldsymbol{\eta}_t$ is $n \times 1$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Lambda}$ are both diagonal (with elements $\omega_i^2 \geq 0$ and $\lambda_i^2 > 0$, respectively), and $\boldsymbol{\Phi}_0$ is lower triangular with ones on the diagonal. We shall refer to this as the *expanded VMA form*; the autocovariances implied by this decomposition are

$$\mathring{\boldsymbol{\Gamma}}_j = \sum_{l=j}^{q} \boldsymbol{\Phi}_l \boldsymbol{\Omega} \boldsymbol{\Phi}_{l-j}' + \mathbb{1}(j=0)\boldsymbol{\Lambda}, \qquad j = 0, \dots, q, \tag{5}$$

and the mapping between $(\boldsymbol{\Theta}_0, \dots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma})$ and $(\boldsymbol{\Phi}_0, \dots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}, \boldsymbol{\Lambda})$ is established by setting $\boldsymbol{\Gamma}_j = \mathring{\boldsymbol{\Gamma}}_j$, i.e.,

$$\sum_{l=j}^{q} \boldsymbol{\Theta}_l \boldsymbol{\Sigma} \boldsymbol{\Theta}_{l-j}' = \sum_{l=j}^{q} \boldsymbol{\Phi}_l \boldsymbol{\Omega} \boldsymbol{\Phi}_{l-j}' + \mathbb{1}(j=0)\boldsymbol{\Lambda}, \quad \text{for all } j = 0, \dots, q. \tag{6}$$

The theoretical justification for decomposing $\mathbf{u}_t$ this way is as follows. Peiris (1988, Theorem 2) proved that the sum of any two independent VMA processes yields a VMA process. The following theorem shows that any VMA process $\mathbf{u}_t = \boldsymbol{\Theta}(L)\boldsymbol{\varepsilon}_t$, satisfying $\det \boldsymbol{\Theta}(z) \neq 0$ for all $|z| = 1$, can be written in the form (4).

Theorem 1. *Every VMA(q) process with finite first and second-order moments and no roots on the unit circle can be expressed in the expanded VMA form.*

Proof. Appendix A.2. □

We make two observations regarding the expanded form. First, it is possible that some (although not all) $\omega_i^2 = 0$, which corresponds to $f_{i,t} = 0$. Suppose that $\omega_1^2 > 0, \ldots, \omega_{n_1}^2 > 0$ and $\omega_{n_1+1}^2 = \cdots = \omega_n^2 = 0$. If $n_1 \ll n$, then (4) reduces to a *latent factor model*, where $f_{1,t}, \ldots, f_{n_1,t}$ can be interpreted as n_1 latent factors (see Chan, Eisenstat, and Koop, 2016, Section 3.4, for more details).

In the general case, the decomposition may be regarded as a projection of $\mathbf{u}_t$ onto some closed linear subspace spanned by $\mathbf{f}_t, \mathbf{f}_{t-1}, \ldots$; consequently, there is no sensible interpretation to be attached to $\mathbf{f}_t, \mathbf{f}_{t-1}, \ldots$, in lieu of a specific economic theory. In fact, while Theorem 1 ensures that most VMAs of interest admit an expanded form representation, such representations are generally not unique. In the following subsection, we address this feature and establish useful properties of the expanded form that make it suitable for Bayesian analysis of VMA processes. For the remainder of the paper, we focus on the case where $\omega_i^2 > 0$ for all $i = 1, \ldots, n$.

2.2 Identification in the Expanded VMA Form

Suppose that for estimation purposes, the VMA(q) in (1) is specified with $\mathbf{\Sigma}$ unrestricted and enough constraints on $\mathbf{\Theta}_0, \ldots, \mathbf{\Theta}_q$ such that there are altogether $m = qn^2 + 0.5n(n+1)$ free parameters, which are uniquely identified (i.e., are uniquely recovered from $\mathbf{\Gamma}_0, \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_q$). It is clear that relative to this specification, there are n additional parameters in the corresponding expanded form. Thus, to estimate the expanded form in practice, we must consider the "identification problem" generated by the expansion.

Metaxoglou and Smith (2007) deal with this by fixing the elements of $\mathbf{\Lambda}$ to pre-determined values and estimate the remaining coefficients as free parameters. However, such a strategy might lead to mis-specification (e.g., it automatically imposes arbitrary lower bounds on the process variance). In fact, there seems to be no reasonable approach to "fixing parameters" in this context. Fortunately, it is not necessary either.

To clarify the basic ideas, consider the simplest case of a fundamental MA(1). The mapping between the two forms is defined by

$$\sigma^2(1+\theta^2) = \omega^2(1+\phi^2) + \lambda^2,$$
$$\sigma^2\theta = \omega^2\phi.$$

Note further that in this case, we have $\sigma^2 > 0$, $\omega^2 > 0$, $\lambda^2 > 0$ and $-1 < \theta < 1$. It is easy to show, however, that these equalities and inequalities jointly imply

$$0 < \lambda^2 < \sigma^2(1 - |\theta|)^2. \tag{7}$$

Since θ, σ^2 are uniquely identified from data, λ^2 is always bounded within a finite interval—this interval is largest when $\theta = 0$ (and the model reduces to white noise) and shrinks towards zero as $\theta \to \pm 1$.

This fact has several important implications. Specifically, there always exists (except for the extreme case where $\theta = \pm 1$) some nonzero (positive) λ^2 such that any θ, σ^2 can be recovered from $\phi, \omega^2, \lambda^2$. However, given a particular value of λ^2, the reverse mapping from θ, σ^2 to ϕ, ω^2 is not well defined (i.e., it does not exist for all values of $\sigma^2 > 0$, $-1 < \theta < 1$). On the other hand, while many combinations of $\phi, \omega^2, \lambda^2$ will generally map to the same θ, σ^2, only values of λ^2 satisfying (7) are admissible. Two implications follow:

1. arbitrarily fixing λ^2 at a particular value may lead to mis-specification;
2. $\phi, \omega^2, \lambda^2$ are all partially identified in the expanded form.

The first point demonstrates why a strategy such as the one employed by Metaxoglou and Smith (2007) might not be appropriate; the second suggests that leaving $\phi, \omega^2, \lambda^2$ as unrestricted parameters and employing Bayesian methods should work well in this context.

Before discussing further details, note that the above intuition generalizes in a straightforward way. Let $\boldsymbol{\lambda} = (\lambda_1^2, \ldots, \lambda_n^2)'$ and $\boldsymbol{\varphi}$ be the $m \times 1$ vector of all free parameters in $\boldsymbol{\Phi}_0, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}$. For each $\boldsymbol{\lambda}^* \in \mathbb{R}^n$, define $\mathbb{M}_{\boldsymbol{\lambda}^*} = \{\boldsymbol{\varphi} \in \mathbb{R}^m : \boldsymbol{\lambda} = \boldsymbol{\lambda}^* \text{ and } \boldsymbol{\Gamma}_j = \mathring{\boldsymbol{\Gamma}}_j \text{ for all } j = 0, \ldots, q\}$ and let $\mathbb{L} = \{\boldsymbol{\lambda} \in \mathbb{R}^n : \mathbb{M}_{\boldsymbol{\lambda}} \neq \emptyset\}$. Using this notation, we define the set of all admissible expanded form parameters that correspond to a particular VMA(q) specification as $\mathbb{P} = \{(\boldsymbol{\lambda}, \boldsymbol{\varphi}) \in \mathbb{R}^{m+n} : \boldsymbol{\lambda} \in \mathbb{L} \text{ and } \boldsymbol{\varphi} \in \mathbb{M}_{\boldsymbol{\lambda}}\}$. The following theorem characterizes the expanded form parameter space.

Theorem 2. *Consider a VMA(q) process with VMA(1) representation parameters $\widetilde{\boldsymbol{\Theta}}_0$, $\widetilde{\boldsymbol{\Theta}}_1$, $\widetilde{\boldsymbol{\Sigma}}$ and expanded form parameters $\boldsymbol{\Phi}_0, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}, \boldsymbol{\Lambda}$, where $\omega_i^2 > 0$ for $i = 1, \ldots, n$. For every $\rho \in \mathbb{C}$ satisfying $|\rho| = 1$, define the $qn \times qn$ Hermitian matrix*

$$\mathbf{H}_\rho = \left(\widetilde{\boldsymbol{\Theta}}_0 + \rho \widetilde{\boldsymbol{\Theta}}_1\right) \widetilde{\boldsymbol{\Sigma}} \left(\widetilde{\boldsymbol{\Theta}}_0' + \bar{\rho} \widetilde{\boldsymbol{\Theta}}_1'\right), \tag{8}$$

with eigenvalues (ordered from smallest to largest) $\mu_1(\mathbf{H}_\rho), \ldots, \mu_{qn}(\mathbf{H}_\rho)$. Then, the set $\mathbb{P}$ of all permissible expanded form parameters that correspond to the given VMA(q), is a bounded subset of $\mathbb{R}^{m+n}$ with the properties:

1. *for every $\boldsymbol{\lambda} \in \mathbb{L}$, the r-th largest λ_i^2 is bounded on the interval*

$$0 < \lambda_i^2 \leq \min_{|\rho|=1} (\mu_{qr}(\mathbf{H}_\rho)); \tag{9}$$

2. *for every $\boldsymbol{\lambda} \in \mathbb{L}$, the set $\mathbb{M}_{\boldsymbol{\lambda}}$ is finite.*

Corollary 1. *Let z^* be a root of $\boldsymbol{\Theta}(z)$. Then as $|z^*| \longrightarrow 1$, at least one (possibly all) $\lambda_i^2 \longrightarrow 0$.*

Proof. Appendix A.2. □

In a Bayesian context, Theorem 2 provides guidance for setting priors on the expanded form parameters. Specifically, a Bayesian approach may generally proceed by

1. sampling $\boldsymbol{\Phi}_0, \ldots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}, \boldsymbol{\Lambda}$ using the expanded form;
2. recovering $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$ ex-post from the simulated draws.

It is important to emphasize that identification of $\boldsymbol{\Phi}_0, \ldots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}, \boldsymbol{\Lambda}$ is not necessary to successfully implement Step 1. This is because a well-defined posterior distribution may be obtained even when the likelihood does not uniquely identify the parameters in the model. For example, if all priors are proper, then the posterior is always proper (Poirier, 1995). The same may hold for certain classes of improper priors as well. The key issue in a Bayesian analysis is that the weaker the identification in the likelihood, the more sensitive is the posterior to the prior; in the extreme case where no likelihood identification exists, the posterior will simply be equal to the prior.

In the present context, we are not interested in the posterior distribution of $\boldsymbol{\Phi}_0, \ldots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}$, and $\boldsymbol{\Lambda}$ itself; we only use it as a means to analyze the posterior of identified quantities such as $\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$ or the forecasts $\mathbf{u}_{t+1}, \ldots, \mathbf{u}_{t+h}$. To this end, recall that sampling directly from any posterior distribution $(\widetilde{\boldsymbol{\theta}} \,|\, \mathbf{y})$ and applying the transformation $g : \widetilde{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ to each draw of $\widetilde{\boldsymbol{\theta}}$ automatically yields samples from the posterior $(\boldsymbol{\theta} \,|\, \mathbf{y})$. If sampling is easier from $(\widetilde{\boldsymbol{\theta}} \,|\, \mathbf{y})$ than from $(\boldsymbol{\theta} \,|\, \mathbf{y})$, such a two-step approach provides clear computational gains. Moreover, if restrictions are needed to, say, ensure a uni-modal $(\boldsymbol{\theta} \,|\, \mathbf{y})$, then it makes sense to incorporate such restrictions into the transformation g, and apply them in the post-processing of draws.

The underlying premise of the approach we propose is that sampling from the expanded form is computationally easier than sampling from the standard VMA posterior. Indeed, sampling from an unidentified (or partially identified) parameter space, then transforming to draws from an identified parameter space is a common strategy employed by Bayesians to improve computation (examples include Meng and van Dyk, 1999; Liu and Wu, 1999; Gustafson, 2005; Imai and van Dyk, 2005; Ghosh and Dunson, 2009; Koop, León-González, and Strachan, 2010, 2012). The primary computational advantage of the expanded VMA form, is that it can be cast as a *linear* state space model, for which efficient MCMC algorithms already exist (e.g., Durbin and Koopman, 2002; Chan and Jeliazkov, 2009). Moreover, as detailed in Subsection 2.3 and Appendix A.1, it is straightforward to implement restrictions on the roots of $\boldsymbol{\Theta}(L)$ in the post-processing of expanded form parameter draws (Step 2), when such restrictions are necessary for identification.

Following this reasoning, we propose to assign priors to each $\lambda_1^2, \ldots, \lambda_n^2$ instead of fixing them to specific values. However, priors on expanded form parameters affect both the efficiency of the sampling algorithm and the posterior of the VMA parameters of interest through the implied priors on the latter, so it is important to do this prudently. One implication of Theorem 2 is that priors on $\lambda_1^2, \ldots, \lambda_n^2$ correspond to priors on the roots of $\det \boldsymbol{\Theta}(L)$: lower prior probabilities assigned to small values of λ_i^2 imply higher probabilities of restricting the roots to be away from unity. On the other hand, the upper bound on λ_i^2 is identified from the data, so the tail of the prior plays a minor role in determining the posterior of $\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$.

Consequently, a sensible prior on λ_i^2 may be formulated using the inverse-gamma distribution as

$$\lambda_i^2 \sim \mathcal{IG}(\nu_{\lambda,0}, S_{\lambda,0}),$$

with $\nu_{\lambda,0}, S_{\lambda,0}$ set to low values. Setting a small $S_{\lambda,0}$ ensures that VMA processes with roots close to unity are not excluded *a priori*; setting a low $\nu_{\lambda,0}$ results in a fat-tailed distribution that improves the mixing efficiency of MCMC algorithms based on this specification. In our extensive experimentation (with a variety of data sets and VARMA models) using this prior, we have consistently found that mixing improves as $\nu_{\lambda,0}$ is decreased for any given $S_{\lambda,0}$ and the prior is flattened. We discuss priors on the remaining expanded form parameters in more detail in Section 3.

2.3 Recovering $(\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma})$ from $(\boldsymbol{\Phi}_0, \ldots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}, \boldsymbol{\Lambda})$

For forecasting applications, the expanded VMA form by construction yields identical predictive distributions to the standard VMA form. Therefore, there is no need to recover $\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$, and forecasts $\mathbf{u}_{t+1}, \ldots, \mathbf{u}_{t+h}$ can be generated directly from draws of expanded form parameters.

However, in many applications—particularly those focused on analyzing impulse responses—the posterior of $\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$ is of primary interest. Fortunately, it is straightforward to recover draws of $\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$ from draws of $\boldsymbol{\Phi}_0, \ldots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}$ and $\boldsymbol{\Lambda}$. For clarity and to highlight a key computational advantage of using the expanded form, we describe the procedure assuming that $\boldsymbol{\varepsilon}_t$ is Gaussian.[2]

A well-known feature of VMA models with Gaussian errors is that *fundamental* and *non-fundamental* specifications are observationally equivalent. In practice, it is common to estimate the fundamental one (Canova, 2007, Chapter 4). However, many theoretical macroeconomic models (e.g. Yang, 2005; Leeper, Walker, and Yang, 2008) imply

[2]When the white noise process $\{\boldsymbol{\varepsilon}_t\}$ is Gaussian, condition (6) is *necessary and sufficient* for any given expanded form to be an observationally equivalent representation of the VMA(q) process. This is because in the Gaussian case, the spectral density of $\{\mathbf{u}_t\}$ completely characterizes all stochastic properties of the series. When $\{\boldsymbol{\varepsilon}_t\}$ is non-Gaussian, however, condition (6) is *necessary* but *not sufficient* in the sense that a given expanded form satisfying (6) may not preserve the *generalized spectral density* of $\{\mathbf{u}_t\}$, which further accounts for *non-linear dependence* in the series (Hamilton and Lin, 1996).

that impulse responses should be computed from non-fundamental representations, while Lippi and Reichlin (1994) argue that economic theory is rarely informative about a particular non-fundamental representation (even when strong justification exists for non-fundamentalness in general), and therefore, impulse responses from the fundamental and all *basic non-fundamental* representations should be reported.[3]

With respect to the expanded form, note that $\boldsymbol{\varepsilon}_t$ is a Gaussian white noise process *if and only if* $\mathbf{f}_t$ and $\boldsymbol{\eta}_t$ are Gaussian white noise processes. This makes the Gaussian framework particularly convenient for implementing algorithms based on the expanded form. Moreover, an important consequence of sampling from the expanded form is that draws from all fundamental and basic non-fundamental VMA representations are easily obtained from draws of expanded form parameters.

In particular, assume that $\boldsymbol{\Theta}_0$ is known, $\boldsymbol{\Sigma}$ is unrestricted, and we have obtained draws of $\boldsymbol{\Phi}_0, \dots \boldsymbol{\Phi}_q, \boldsymbol{\Omega}$, and $\boldsymbol{\Lambda}$.[4] We proceed to recover $\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$ (given $\boldsymbol{\Theta}_0$) by appealing to the corresponding VMA(1) representation, which yields the system of equations:

$$\widetilde{\boldsymbol{\Gamma}}_0 = \widetilde{\boldsymbol{\Theta}}_0\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Theta}}_0' + \widetilde{\boldsymbol{\Theta}}_1\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Theta}}_1' = \widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\Theta}}_1\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}_1', \tag{10}$$

$$\widetilde{\boldsymbol{\Gamma}}_1 = \widetilde{\boldsymbol{\Theta}}_1\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Theta}}_0' = \widehat{\boldsymbol{\Theta}}_1\widehat{\boldsymbol{\Sigma}}, \tag{11}$$

where $\widehat{\boldsymbol{\Theta}}_1 = \widetilde{\boldsymbol{\Theta}}_1\widetilde{\boldsymbol{\Theta}}_0^{-1}$, $\widehat{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Theta}}_0\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Theta}}_0'$, and $\widetilde{\boldsymbol{\Gamma}}_0$, $\widetilde{\boldsymbol{\Gamma}}_1$ are computed from $\boldsymbol{\Phi}_0, \dots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}$ and $\boldsymbol{\Lambda}$ by setting $\boldsymbol{\Gamma}_j = \mathring{\boldsymbol{\Gamma}}_j$ for $j = 0, \dots, q$. Left-multiplying (10) by $\widehat{\boldsymbol{\Theta}}_1$ and substituting (11) into (10) yields the matrix quadratic equation

$$\widehat{\boldsymbol{\Theta}}_1^2\widetilde{\boldsymbol{\Gamma}}_1' - \widehat{\boldsymbol{\Theta}}_1\widetilde{\boldsymbol{\Gamma}}_0 + \widetilde{\boldsymbol{\Gamma}}_1 = 0. \tag{12}$$

In Appendix A.1 we provide the computational details of solving the matrix quadratic equation (12). To summarize, the algorithm to obtain draws of $\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$ (given $\boldsymbol{\Theta}_0$) from draws of $\boldsymbol{\Phi}_0, \dots \boldsymbol{\Phi}_q, \boldsymbol{\Omega}$, and $\boldsymbol{\Lambda}$ consists of the following four steps:

1. Compute $\boldsymbol{\Gamma}_0, \dots, \boldsymbol{\Gamma}_q$ from draws of $\boldsymbol{\Phi}_0, \dots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}, \boldsymbol{\Lambda}$.
2. Construct $\widetilde{\boldsymbol{\Gamma}}_0$ and $\widetilde{\boldsymbol{\Gamma}}_1$ according to (3).
3. Compute $\widehat{\boldsymbol{\Theta}}_1$, the solution of (12), and the corresponding $\widehat{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Gamma}}_0 - \widehat{\boldsymbol{\Theta}}_1\widetilde{\boldsymbol{\Gamma}}_1'$. The roots of $\boldsymbol{\Theta}(L)$ are selected as a byproduct of this step (see Appendix A.1 for details).

[3] Lippi and Reichlin (1994) define *basic non-fundamental* representations as all non-fundamental representations that are obtained from the fundamental one by "flipping" one or more of its MA roots. An important aspect of this classification is that while in a general VARMA(p, q) other non-fundamental representations are possible, Lippi and Reichlin (1994) show that only basic non-fundamental representations preserve the AR and MA orders (p and q). Based on this, they argue that only basic non-fundamental representations should be considered in empirical work.

[4] In practice, some restrictions are needed on $\boldsymbol{\Theta}_0, \dots \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$, to achieve identification. Typical applications employ the restriction $\boldsymbol{\Theta}_0 = \mathbf{I}_n$. In VARMAs specified with the echelon form or scalar component models (Athanasopoulos, Poskitt, and Vahid, 2012), the restriction $\boldsymbol{\Theta}_0 = \mathbf{B}_0$ is used, where $\mathbf{B}_0$ are coefficients in the conditional mean. Our approach readily accommodates these cases as well any other specification where $\boldsymbol{\Theta}_0$ is determined outside of the moving average expansion and does not rely explicitly on expanded form parameters.

4. Recover $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_q$ from the first n columns of $\widehat{\boldsymbol{\Theta}}_1$ and $\boldsymbol{\Sigma}$ from the bottom-right $n \times n$ block of $\widehat{\boldsymbol{\Sigma}}$.

Once the VMA parameters ($\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_q$, $\boldsymbol{\Sigma}$) are recovered, the reduced-formed errors $\{\boldsymbol{\varepsilon}_t\}$ in the original parameterization can be computed from (1).

2.4 Extension to the VARMA

It is straightforward to generalize the above setup to VARMAs. Specifically we add p autoregressive components to $\mathbf{u}_t$ in (1) to obtain

$$\mathbf{y}_t = \sum_{j=1}^{p} \mathbf{A}_j \mathbf{y}_{t-j} + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{\varepsilon}_{t-j} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{WN}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{13}$$

where the intercept is suppressed for notational convenience. Following the same procedure as before, we derive the *expanded VARMA form* by reparameterizing $\mathbf{u}_t$ such that

$$\mathbf{y}_t = \sum_{j=1}^{p} \mathbf{A}_j \mathbf{y}_{t-j} + \sum_{j=0}^{q} \boldsymbol{\Phi}_j \mathbf{f}_{t-j} + \boldsymbol{\eta}_t, \qquad \begin{pmatrix} \mathbf{f}_t \\ \boldsymbol{\eta}_t \end{pmatrix} \sim \mathcal{WN}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda} \end{pmatrix} \right). \tag{14}$$

Since we can easily recover $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$ from $\boldsymbol{\Phi}_0, \ldots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}, \boldsymbol{\Lambda}$, it is clear that estimating (14) is sufficient to estimate (13).

It is important to point out, however, that introducing autoregressive components in this context leads to new complications. Specifically, it is well known that a VARMA(p, q) specified as in (13) may not be identified (see, for example, Lütkepohl, 2005). A number of ways have been proposed in the literature to deal with this. For example, one may impose the canonical echelon form by reformulating (13) as

$$\mathbf{B}_0 \mathbf{y}_t = \sum_{j=1}^{p} \mathbf{B}_j \mathbf{y}_{t-j} + \sum_{j=1}^{p} \boldsymbol{\Theta}_j^* \boldsymbol{\varepsilon}_{t-j} + \mathbf{B}_0 \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{WN}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{15}$$

where $\mathbf{B}_0$ is lower-triangular with ones on the diagonal, $\mathbf{B}_j = \mathbf{B}_0 \mathbf{A}_j$ and $\boldsymbol{\Theta}_j^* = \mathbf{B}_0 \boldsymbol{\Theta}_j$. Conditional on the system's Kronecker indices $\kappa_1, \ldots, \kappa_n$, with $0 \leq \kappa_i \leq p$, (15) can be estimated by imposing exclusion restrictions on the coefficients in $\mathbf{B}_0$, $\{\mathbf{B}_j\}$ and $\{\boldsymbol{\Theta}_j^*\}$.

Observe that conditional on the Kronecker indices, it is straightforward to impose echelon form restrictions on the expanded VARMA as well, by rewriting (14) as

$$\mathbf{B}_0 \mathbf{y}_t = \sum_{j=1}^{p} \mathbf{B}_j \mathbf{y}_{t-j} + \sum_{j=0}^{p} \boldsymbol{\Phi}_j \mathbf{f}_{t-j} + \boldsymbol{\eta}_t, \qquad \begin{pmatrix} \mathbf{f}_t \\ \boldsymbol{\eta}_t \end{pmatrix} \sim \mathcal{WN}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda} \end{pmatrix} \right). \tag{16}$$

Clearly, exclusion restrictions can be easily imposed on the elements of $\mathbf{B}_0, \{\mathbf{B}_j\}$ and $\{\boldsymbol{\Phi}_j\}$ for sampling purposes (see Appendix A.3 for more details). More importantly,

the echelon form requires that only full-row restrictions are imposed on $\{\boldsymbol{\Theta}_j^*\}$. Because restricting any row of $\boldsymbol{\Phi}_j$ to zero will correspond to restricting the same row of $\boldsymbol{\Theta}_j^*$ to zero under (6), we can once again work directly with (16) and recover the parameters of (15) *ex post*.

In practice, of course, Kronecker indices will be unknown, and in our related paper (Chan, Eisenstat, and Koop, 2016), we specify priors for Kronecker indices and construct efficient sampling procedures—based on the expanded VARMA form developed in the present paper—to jointly estimate Kronecker indices and model parameters.[5] For the remainder of the present paper, however, we will rely on the identifying assumption that the concatenated matrix $[\mathbf{A}_p : \boldsymbol{\Theta}_q]$ has full rank n (e.g., Hannan, 1976).

This assumption will generally hold when q and n are relatively small, and has the advantage of being much simpler than specifying the Echelon form. The drawback of this approach is that the resulting representation is not canonical in the sense that not all VARMA processes satisfy this assumption (Lütkepohl and Poskitt, 1996). In our application, however, the main interest is to assess whether adding moving average terms to a time-varying parameter VAR with stochastic volatility improves forecasts of macroeconomic variables. In the time-varying parameter context, where typically only small systems are considered, we find the above identification strategy to be suitable, and as further discussed in Section 4, adding even a small number of moving average components may lead to substantial gains in forecast accuracy.

Identification based on the full rank of $[\mathbf{A}_p : \boldsymbol{\Theta}_q]$ assumption also becomes more difficult to justify as as n increases. Therefore, for larger systems one may wish to consider canonical specifications and follow the approach in Chan, Eisenstat, and Koop (2016). A key point is that regardless of the scheme used to uniquely identify $\mathbf{A}(L)$ and $\boldsymbol{\Theta}(L)$, the expanded form representation provides a convenient framework for developing sampling algorithms in VARMA specifications.

3 Estimation of VARMA with TVP and SV

In this section, we first consider a general VARMA with time-varying coefficients and stochastic volatility. We then introduce the conjugate priors for the model parameters, followed by a discussion of an efficient Gibbs sampler. We note that throughout, the analysis is performed conditional on the initial observations $\mathbf{y}_0, \dots, \mathbf{y}_{1-p}$ and assuming the initial factors $\mathbf{f}_{1-q} = \cdots = \mathbf{f}_0 = \mathbf{0}$. One can extend the posterior sampler to the case where the initial observations or factors are modeled explicitly. Moreover, we will assume that the VARMA is specified with an intercept term $\boldsymbol{\mu}$. Again, the ensuing algorithm is easily extended to include additional exogenous variables.

As highlighted previously, the key advantage of working directly with the expanded

[5]Note that in Chan, Eisenstat, and Koop (2016), the focus is on large, constant parameter VARMAs, whereas the main concern of this paper is small, time-varying parameter VARMAs.

VARMA form is that it is conditionally linear, and therefore, leads to straightforward computation. This in turn opens the door to a wealth of extensions that have already been well developed for linear models, but have thus far been inaccessible for even the simplest of VMA specifications. Our particular interest is to enhance the basic VARMA(p, q) with two extensions particularly relevant for empirical macroeconomic applications: stochastic volatility and time-varying parameters.

To that end, we extend (14) to allow the VMA coefficients $\mathbf{\Phi}_0, \ldots, \mathbf{\Phi}_q$ to be time-varying:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{\Phi}_{0,t}\mathbf{f}_t + \mathbf{\Phi}_{1,t}\mathbf{f}_{t-1} + \cdots + \mathbf{\Phi}_{q,t}\mathbf{f}_{t-q} + \boldsymbol{\eta}_t, \tag{17}$$

where $\mathbf{X}_t = \mathbf{I}_n \otimes (1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})$, $\boldsymbol{\beta} = \text{vec}((\boldsymbol{\mu}, \mathbf{A}_1, \ldots, \mathbf{A}_p)')$, $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$, and $\mathbf{\Lambda}$ is a diagonal matrix.

Let $\boldsymbol{\phi}_{i,t}$ denote the vector of free parameters in the i-th row of $(\mathbf{\Phi}_{0,t}, \mathbf{\Phi}_{1,t}, \ldots, \mathbf{\Phi}_{q,t})$. Note that the dimension of $\boldsymbol{\phi}_{i,t}$ is k_i with $k_i = i - 1 + nq$. Then, consider the transition equation

$$\boldsymbol{\phi}_{i,t} = \boldsymbol{\phi}_{i,t-1} + \boldsymbol{\xi}_{i,t}, \tag{18}$$

where $\boldsymbol{\xi}_{i,t} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}_{\phi_i})$ for $i = 1, \ldots, n, t = 2, \ldots, T$ with $\mathbf{\Psi}_{\phi_i} = \text{diag}(\psi^2_{\phi,i,1}, \ldots, \psi^2_{\phi,i,k_i})$. For later reference, stack $\boldsymbol{\psi}^2_\phi = (\psi^2_{\phi,1,1}, \ldots, \psi^2_{\phi,n,k_n})'$. The initial conditions are specified as $\boldsymbol{\phi}_{i,1} \sim \mathcal{N}(\boldsymbol{\phi}_{i,0}, \mathbf{\Psi}_{\phi_0})$, where $\boldsymbol{\phi}_{i,0}$ and $\mathbf{\Psi}_{\phi_0}$ are known constant matrices.

Next, we incorporate stochastic volatility into the model by allowing the latent factors to have time-varying volatilities $\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_t)$, where $\mathbf{\Omega}_t = \text{diag}(\text{e}^{h_{1,t}}, \ldots, \text{e}^{h_{n,t}})$. Then, each of the log-volatility follows an independent random walk:

$$h_{i,t} = h_{i,t-1} + \zeta_{i,t}, \tag{19}$$

where $\zeta_{i,t} \sim \mathcal{N}(0, \psi^2_{h,i})$ for $i = 1, \ldots, n, t = 2, \ldots, T$. The log-volatilities are initialized with $h_{i,1} \sim \mathcal{N}(h_{i,0}, V_{h_{i,0}})$, where $h_{i,0}$ and $V_{h_{i,0}}$ are known constants. For notational convenience, let $\mathbf{h}_t = (h_{1,t}, \ldots, h_{n,t})'$, $\mathbf{h} = (\mathbf{h}'_1, \ldots, \mathbf{h}_T)'$ and $\boldsymbol{\psi}^2_h = (\psi^2_{h,1}, \ldots, \psi^2_{h,n})'$. Note that allowing for both time-varying $\mathbf{\Phi}_{0,t}, \ldots, \mathbf{\Phi}_{q,t}$ and $\mathbf{\Omega}_t$ will correspond to fully time-varying $\mathbf{\Theta}_{1,t}, \ldots, \mathbf{\Theta}_{j,t}$ and $\mathbf{\Sigma}_t$, while normal distributions assigned to $\mathbf{f}_t$ and $\boldsymbol{\eta}_t$ imply $\boldsymbol{\varepsilon}_t$ is distributed conditionally normal as well.

To facilitate estimation, stack the observations in (17) over t:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{\Phi}\mathbf{f} + \boldsymbol{\eta}, \tag{20}$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_T \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} \mathbf{f}_1 \\ \vdots \\ \mathbf{f}_T \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_T \end{pmatrix},$$

and $\mathbf{\Phi}$ is a a $Tn \times Tn$ lower triangular matrix with $\mathbf{\Phi}_{0,1}, \ldots, \mathbf{\Phi}_{0,T}$ on the main diagonal block, $\mathbf{\Phi}_{1,2}, \ldots, \mathbf{\Phi}_{1,T}$ on first lower diagonal block, $\mathbf{\Phi}_{2,3}, \ldots, \mathbf{\Phi}_{2,T}$ on second lower diagonal

block, and so forth. For example, for $q = 2$, we have

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{\Phi}_{0,1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{\Phi}_{1,2} & \mathbf{\Phi}_{0,2} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{\Phi}_{2,3} & \mathbf{\Phi}_{1,3} & \mathbf{\Phi}_{0,3} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_{2,4} & \mathbf{\Phi}_{1,4} & \mathbf{\Phi}_{0,4} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Phi}_{2,T} & \mathbf{\Phi}_{1,T} & \mathbf{\Phi}_{0,T} \end{pmatrix}.$$

Note that in general $\mathbf{\Phi}$ is a band $Tn \times Tn$ matrix—i.e., its nonzero elements are confined in a narrow band along the main diagonal—that contains at most

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

nonzero elements, which grows *linearly* in T and is substantially less than the total $(Tn)^2$ elements for typical applications where $T \gg q$. This special structure can be exploited to speed up computation, e.g., by using block-banded or sparse matrix algorithms (see, e.g., Kroese, Taimre, and Botev, 2011, p. 220).

To complete the model specification, we assume independent priors for $\boldsymbol{\beta}$, $\mathbf{\Lambda}$, $\boldsymbol{\psi}_\phi^2$ and $\boldsymbol{\psi}_h^2$ as follows. For $\boldsymbol{\beta}$, we consider the multivariate normal prior $\mathcal{N}(\boldsymbol{\beta}_0, \mathbf{V}_{\boldsymbol{\beta}})$. For $\boldsymbol{\psi}_\phi^2$, $\boldsymbol{\psi}_h^2$ and the diagonal elements of $\mathbf{\Lambda} = \text{diag}(\lambda_1^2, \ldots, \lambda_n^2)$, we assume the following priors:

$$\lambda_i^2 \sim \mathcal{IG}(\nu_{\lambda,0}, S_{\lambda,0}), \quad \psi_{\phi,i,j}^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2S_{\phi,0}}\right), \quad \psi_{h,i}^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2S_{h,0}}\right),$$

where $\mathcal{G}$ and $\mathcal{IG}$ denote the gamma and the inverse-gamma distributions respectively. Recall that the parameters $\lambda_1^2, \ldots, \lambda_n^2$ are introduced to facilitate computation—for that purpose we will set the degree of freedom parameter $\nu_{\lambda,0}$ to be small. Following Frühwirth-Schnatter and Wagner (2010), we assume gamma priors on the error variances of the time-varying parameters for two reasons. First, compared to the conventional inverse-gamma prior, a gamma prior has more mass concentrated around small values. Hence, this prior provides shrinkage—*a priori* it favors the more parsimonious constant-coefficient model. Second, as shown in Frühwirth-Schnatter and Wagner (2010), the posterior results under this prior are insensitive to the values of the hyperparameters. In our application, this gamma prior works well. Alternatively, hierarchical priors such as those in Korobilis (2014) can also be considered.

Our approach to constructing the sampling algorithm is based on treating (20) as a latent factor model. Specifically, the Gibbs sampler proceeds by sequentially drawing from

1. $p(\boldsymbol{\beta}, \mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\phi}, \mathbf{h}, \boldsymbol{\psi}_\phi^2, \boldsymbol{\psi}_h^2, \mathbf{\Lambda}) = p(\boldsymbol{\beta} \,|\, \mathbf{y}, \boldsymbol{\phi}, \mathbf{h}, \mathbf{\Lambda}) p(\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{h}, \mathbf{\Lambda})$;
2. $p(\boldsymbol{\phi} \,|\, \mathbf{y}, \boldsymbol{\beta}, \mathbf{f}, \mathbf{h}, \boldsymbol{\psi}_\phi^2, \boldsymbol{\psi}_h^2, \mathbf{\Lambda})$;
3. $p(\mathbf{h} \,|\, \mathbf{y}, \boldsymbol{\beta}, \mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi}_\phi^2, \boldsymbol{\psi}_h^2, \mathbf{\Lambda})$;

4. $p(\boldsymbol{\Lambda}, \boldsymbol{\psi}_\phi^2, \boldsymbol{\psi}_h^2 \,|\, \mathbf{y}, \boldsymbol{\beta}, \mathbf{f}, \mathbf{h}, \boldsymbol{\phi}) = p(\boldsymbol{\Lambda} \,|\, \mathbf{y}, \boldsymbol{\beta}, \mathbf{f}, \boldsymbol{\phi}) p(\boldsymbol{\psi}_h^2 \,|\, \mathbf{h}) p(\boldsymbol{\psi}_\phi^2 \,|\, \boldsymbol{\phi})$.

We initialize the model parameters using values that are consistent with a constant coefficients VAR. Specifically, let $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{S}} = (\widehat{s}_{ij})$ be the least squares estimates of the VAR coefficients and the covariance matrix from a VAR(p). The Gibbs sampler is then initialized by setting $\boldsymbol{\beta} = \widehat{\mathbf{b}}$, $\boldsymbol{\phi} = \mathbf{0}$, $h_{i,1} = \cdots = h_{i,T} = \log \widehat{s}_{ii}$, for $i = 1, \ldots, n$, and $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{O}_h)$, where $\mathbf{O}_h = \text{diag}(\text{e}^{h_{1,1}}, \ldots, \text{e}^{h_{n,1}}, \ldots, \text{e}^{h_{1,T}}, \ldots, \text{e}^{h_{n,T}})$. Finally, $\boldsymbol{\Lambda}$, $\boldsymbol{\psi}_\phi^2$, and $\boldsymbol{\psi}_h^2$ are initialized by implementing Step 4 of the Gibbs sampler.

Next, we give some details of the implementation of the Gibbs sampler above. We focus on Step 1; Step 2 to Step 4 are standard, and we leave the details to the Appendix. Although it might be straightforward to sample $\boldsymbol{\beta}$ and $\mathbf{f}$ separately by simulating $\boldsymbol{\beta}$ given $\mathbf{f}$ followed by drawing $\mathbf{f}$ given $\boldsymbol{\beta}$, such an approach would potentially induce high autocorrelation and slow mixing in the constructed Markov chain as $\boldsymbol{\beta}$ and $\mathbf{f}$ enter (20) additively. Instead, we aim to sample $\boldsymbol{\beta}$ and $\mathbf{f}$ jointly—by first drawing $\boldsymbol{\beta}$ marginally of $\mathbf{f}$, followed by drawing $\mathbf{f}$ given $\boldsymbol{\beta}$ and other model parameters.

To implement the first part, recall that $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{O}_h)$. By integrating out $\mathbf{f}$, the joint density of $\mathbf{y}$ marginal of $\mathbf{f}$ is given by

$$(\mathbf{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{h}, \boldsymbol{\Lambda}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{S}_\mathbf{y}),$$

where $\mathbf{S}_\mathbf{y} = \mathbf{I}_T \otimes \boldsymbol{\Lambda} + \boldsymbol{\Phi}\mathbf{O}_h\boldsymbol{\Phi}'$. Using standard results from linear regression (see, e.g., Kroese and Chan, 2014, p. 239-240), we have

$$(\boldsymbol{\beta} \,|\, \mathbf{y}, \boldsymbol{\phi}, \mathbf{h}, \boldsymbol{\Lambda}) \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, \mathbf{D}_{\boldsymbol{\beta}}),$$

where

$$\mathbf{D}_{\boldsymbol{\beta}} = \left(\mathbf{V}_{\boldsymbol{\beta}}^{-1} + \mathbf{X}'\mathbf{S}_\mathbf{y}^{-1}\mathbf{X}\right)^{-1}, \quad \widehat{\boldsymbol{\beta}} = \mathbf{D}_{\boldsymbol{\beta}}\left(\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0 + \mathbf{X}'\mathbf{S}_\mathbf{y}^{-1}\mathbf{y}\right).$$

Since both $\boldsymbol{\Lambda}$ and $\mathbf{O}_h$ are diagonal matrices and $\boldsymbol{\Phi}$ is a lower triangular matrix, the covariance matrix $\mathbf{S}_\mathbf{y}$ is a band matrix. Consequently, we can exploit this feature to speed up computations. Specifically, to compute $\mathbf{X}'\mathbf{S}_\mathbf{y}^{-1}\mathbf{X}$ or $\mathbf{X}'\mathbf{S}_\mathbf{y}^{-1}\mathbf{y}$, one needs not obtain the $Tn \times Tn$ matrix $\mathbf{S}_\mathbf{y}^{-1}$—this would involve $\mathcal{O}(T^3)$ operations. Instead, we obtain $\mathbf{X}'\mathbf{S}_\mathbf{y}^{-1}\mathbf{y}$ by first solving the system $\mathbf{S}_\mathbf{y}\mathbf{z} = \mathbf{y}$ for $\mathbf{z}$, which can be done in $\mathcal{O}(T)$ operations. The solution is $\mathbf{z} = \mathbf{S}_\mathbf{y}^{-1}\mathbf{y}$ and we return $\mathbf{X}'\mathbf{z} = \mathbf{X}'\mathbf{S}_\mathbf{y}^{-1}\mathbf{y}$, which is the desired quantity. Similarly, $\mathbf{X}'\mathbf{S}_\mathbf{y}^{-1}\mathbf{X}$ can be computed quickly without inverting any big matrices.

Next, we sample all the latent factors $\mathbf{f}$ jointly. Note that even though *a priori* the latent factors are independent, they are no longer independent given $\mathbf{y}$. As such, sampling each $\mathbf{f}_t$ sequentially would potentially induce high autocorrelation and slow mixing in the Markov chain. One could sample $\mathbf{f}$ using Kalman filter-based algorithms, but they would involve redefining the states so that only the state at time t enters the measurement equation at time t. As such, each (new) state vector would be of much higher dimension, which in turn results in slower algorithms. Instead, we avoid the Kalman filter and instead implement the precision-based sampler developed in Chan and Jeliazkov (2009) to sample

the latent factors jointly. To that end, recall that *a priori* $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{O}_h)$. Using (20) and standard linear regression results again, we have

$$(\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{h}, \boldsymbol{\Lambda}) \sim \mathcal{N}(\widehat{\mathbf{f}}, \mathbf{K}_{\mathbf{f}}^{-1}),$$

where

$$\mathbf{K}_{\mathbf{f}} = \mathbf{O}_h^{-1} + \boldsymbol{\Phi}'(\mathbf{I}_T \otimes \boldsymbol{\Lambda}^{-1})\boldsymbol{\Phi}, \quad \widehat{\mathbf{f}} = \mathbf{K}_{\mathbf{f}}^{-1}\boldsymbol{\Phi}'(\mathbf{I}_T \otimes \boldsymbol{\Lambda}^{-1})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The challenge, of course, is that the covariance matrix $\mathbf{K}_{\mathbf{f}}^{-1}$ is a $Tn \times Tn$ full matrix, and sampling $(\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{h}, \boldsymbol{\Lambda})$ using brute force is time-consuming. However, the precision matrix $\mathbf{K}_{\mathbf{f}}$ is banded (recall that both $\boldsymbol{\Lambda}^{-1}$ and $\mathbf{O}_h^{-1}$ are diagonal, and $\boldsymbol{\Phi}$ is a band matrix). Again, this feature can be exploited to speed up computation. As before, we first obtain $\widehat{\mathbf{f}}$ by solving

$$\mathbf{K}_{\mathbf{f}}\mathbf{z} = \boldsymbol{\Phi}'(\mathbf{I}_T \otimes \boldsymbol{\Lambda}^{-1})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

for $\mathbf{z}$. Next, obtain the Cholesky decomposition of $\mathbf{K}_{\mathbf{f}}$ such that $\mathbf{C}_{\mathbf{f}}\mathbf{C}_{\mathbf{f}}' = \mathbf{K}_{\mathbf{f}}$. Solve $\mathbf{C}_{\mathbf{f}}'\mathbf{z} = \mathbf{u}$ for $\mathbf{z}$, where $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{Tn})$. Finally, return $\mathbf{f} = \widehat{\mathbf{f}} + \mathbf{z}$, which follows the desired distribution. We refer the readers to Chan and Jeliazkov (2009) for details. The details of Step 2 to Step 4 are given in the appendix.

To give a sense of the speed of the algorithm, we implement it using MATLAB on a desktop with an Intel Core i7-870 @2.93 GHz processor on the dataset in the application with $n = 2$ variables, $T = 215$ observations and $p = 2$ lags. It takes 48 seconds to obtain 10000 posterior draws.

4 Empirical Application

In this section we illustrate the proposed approach and estimation methods with a recursive forecasting exercise that involves US CPI inflation and real GDP growth. These two variables are commonly used in forecasting (e.g., Banbura, Giannone, and Reichlin, 2010; Koop, 2011) and small DSGE models (e.g., An and Schorfheide, 2007). We first outline the set of competing models in Section 4.1, followed by a brief description of the data and the priors. The results of the density forecasting exercise are reported in Section 4.2.

4.1 Competing Models, Data and Prior

The main goal of this forecasting exercise is to illustrate the methodology and investigate how VARMAs and the variants with stochastic volatility compare with standard VARs. We consider four sets of VARMAs: VARMA$(p, 1)$, two versions with different time-varying VMA coefficients and volatility but constant VAR coefficients, and the most flexible version where the VAR coefficients are also time-varying.

More specifically, the VARMA$(p, 1)$ is the same as given in (14). In the first version with time-varying volatility, we allow the latent factors in (14) to have a stochastic

volatility component: $\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_t)$, where $\mathbf{\Omega}_t = \text{diag}(\text{e}^{h_{1,t}}, \ldots, \text{e}^{h_{n,t}})$, and each of the log-volatilities follows an independent random walk as in (19). We call this version VARMA$(p, 1)$-SV1. In the second version, we also allow the matrices $\mathbf{\Phi}_0, \ldots, \mathbf{\Phi}_q$ to be time-varying as specified in (17). This more general version is denoted as VARMA$(p, 1)$-SV2. Finally, we further allow the VAR coefficients $\boldsymbol{\beta} = \text{vec}((\boldsymbol{\mu}, \mathbf{A}_1, \ldots, \mathbf{A}_p)')$ in (17) to be time-varying according to the random walk:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \mathbf{u}_t, \quad \mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}_\beta),$$

where $\mathbf{\Psi}_\beta = \text{diag}(\psi^2_{\beta,1}, \ldots, \psi^2_{\beta,k})$ is diagonal with $k = n(p+1)$. This version is denoted as TVP-VARMA$(p, 1)$-SV2. For comparison we also include standard VAR(p), VAR(p) with stochastic volatility and time-varying parameter VAR(p) with stochastic volatility. These are denoted respectively as VAR(p), VAR(p)-SV, and TVP-VAR(p)-SV.

The data consist of US quarterly CPI inflation and real GDP growth from 1959:Q1 to 2011:Q4. More specifically, given the quarterly real GDP series w_{1t}, we transform it via $y_{1t} = 400 \log(w_{1t}/w_{1,t-1})$ to obtain the growth rate. We perform a similar transformation to the CPI index to get the inflation rate. For easy comparison, we choose broadly similar priors across models. For instance, the priors for the VAR coefficients in VARMA specifications are exactly the same as those of the corresponding VAR.

As discussed in Section 3, we assume the following independent priors: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{V}_{\boldsymbol{\beta}})$, $\boldsymbol{\phi}_i \sim \mathcal{N}(\boldsymbol{\phi}_{0i}, \mathbf{V}_{\boldsymbol{\phi}_i})$, $\omega_i^2 \sim \mathcal{IG}(\nu_{\omega 0}, S_{\omega 0})$ and $\lambda_i^2 \sim \mathcal{IG}(\nu_{\lambda 0}, S_{\lambda 0})$, $i = 1, \ldots, n$.

There is a lot of empirical work that shows using Minnesota-type priors of Doan, Litterman, and Sims (1984); Litterman (1986) for the VAR coefficients improves forecast performance. This suggests that priors that induce shrinkage are crucial in our context. Following this tradition, we set $\boldsymbol{\beta}_0 = \mathbf{0}$ and set the prior covariance $\mathbf{V}_{\boldsymbol{\beta}}$ to be diagonal, where the variances associated with the intercepts are 100 and those corresponding to the VAR coefficients are 1. In other words, this prior induces some shrinkage on the VAR coefficients but not the intercepts.

For the prior on $\boldsymbol{\phi}_i$, we set $\boldsymbol{\phi}_{0i} = \mathbf{0}$ and $\mathbf{V}_{\boldsymbol{\phi}_i}$ to be the identity matrix. By setting the prior mean of $\boldsymbol{\phi}$ to be zero, we effectively shrink a VARMA to a VAR. We choose relatively small values for the degrees of freedom and scale hyperparameters for ω_i^2, which imply large prior variances: $\nu_{\omega 0} = 3$, $S_{\omega 0} = 2$. These values imply $\mathbb{E}\,\omega_i^2 = 1$, $i = 1, \ldots, n$. Finally, for each λ_i^2, we specify the noninformative prior $\nu_{\lambda 0} = 0$, $S_{\lambda 0} = 0.1$.

For VARMAs with stochastic volatility, we need to specify priors on $\psi^2_{h,i}$ and $\psi^2_{\phi,i,j}$. As discussed in Section 3, we assume gamma priors $\psi^2_{h,i} \sim \mathcal{G}(0.5, 0.5/S_{h,0})$ and $\psi^2_{\phi,i,j} \sim \mathcal{G}(0.5, 0.5/S_{\phi,0})$, where $S_{h,0} = 0.01$ and $S_{\phi,0} = 0.001$. These hyperparameters imply $\mathbb{E}\,\psi^2_{h,i} = 0.01$ and $\mathbb{E}\,\psi^2_{\phi,i,j} = 0.001$. For models with time-varying VAR coefficients, we again consider gamma priors for the error variances: $\psi^2_{\beta,i} \sim \mathcal{G}(0.5, 0.5/S_{\beta,0})$ with $S_{\beta,0} = 0.001$, which implies $\mathbb{E}\,\psi^2_{\beta,i} = 0.001$.

4.2 Forecasting Results

To compare the performance of the competing models in producing density forecasts, we consider a recursive out-of-sample forecasting exercise at various forecast horizons as follows. At the t-th iteration, for each of the model we use data up to time t, denoted as $\mathbf{y}_{1:t}$, to construct the joint predictive density $p(\mathbf{y}_{t+k} \,|\, \mathbf{y}_{1:t})$ under the model, and use it as the k-step-ahead density forecast for $\mathbf{y}_{t+k}$. We then expand the sample using data up to time $t+1$, and repeat the whole exercise. We continue this procedure until time $T-k$. At the end of the iterations, we obtain density forecasts under the competing models from 1975Q1 till the end of the sample.

The joint predictive density $p(\mathbf{y}_{t+k} \,|\, \mathbf{y}_{1:t})$ is not available analytically, but it can be estimated using MCMC methods. For VARs and VARMAs, the conditional density of $\mathbf{y}_{t+k}$ given the data and the model parameters—denoted as $p(\mathbf{y}_{t+k} \,|\, \mathbf{y}_{1:t}, \boldsymbol{\theta})$—is Gaussian with known mean vector and covariance matrix. Hence, the predictive density can be estimated by averaging $p(\mathbf{y}_{t+k} \,|\, \mathbf{y}_{1:t}, \boldsymbol{\theta})$ over the MCMC draws of $\boldsymbol{\theta}$. For VARMAs with stochastic volatility, at every MCMC iteration given the model parameters and all the states up to time t, we simulate future log-volatilities from time $t+1$ to $t+k$ using the transition equation. Given these draws, $\mathbf{y}_{t+k}$ has a Gaussian density. Finally, these Gaussian densities are averaged over the MCMC iterations to obtain the joint predictive density $p(\mathbf{y}_{t+k} \,|\, \mathbf{y}_{1:t})$.

To evaluate the quality of the joint density forecast, consider the predictive likelihood $p(\mathbf{y}_{t+k} = \mathbf{y}^{\text{o}}_{t+k} \,|\, \mathbf{y}_{1:t})$, i.e., the joint predictive density of $\mathbf{y}_{t+k}$ evaluated at the observed value $\mathbf{y}^{\text{o}}_{t+k}$. Intuitively, if the actual outcome $\mathbf{y}^{\text{o}}_{t+k}$ is unlikely under the density forecast, the value of the predictive likelihood will be small, and vise versa. We then evaluate the joint density forecasts using the sum of log predictive likelihoods, which is a standard metric in the literature (see, e.g., Clark, 2011; Chan, Koop, Leon-Gonzalez, and Strachan, 2012; Belmonte, Koop, and Korobilis, 2014):

$$\sum_{t=t_0}^{T-k} \log p(\mathbf{y}_{t+k} = \mathbf{y}^{\text{o}}_{t+k} \,|\, \mathbf{y}_{1:t}).$$

This measure can also be viewed as an approximation of the log marginal likelihood; see, e.g., Geweke and Amisano (2011) for a more detailed discussion. In addition to assessing the joint density forecasts, we also evaluate the performance of the models in terms of the forecasts of each individual series. For example, to evaluate the performance for forecasting the i-th component of $\mathbf{y}_{t+k}$, we simply replace the joint density $p(\mathbf{y}_{t+k} = \mathbf{y}^{\text{o}}_{t+k} \,|\, \mathbf{y}_{1:t})$ by the marginal density $p(y_{i,t+k} = y^{\text{o}}_{i,t+k} \,|\, \mathbf{y}_{1:t})$, and report the corresponding sum.

In our forecasting exercise, we present results from the individual models listed in Section 4.1. In addition, we also use a model selection strategy based on the predictive likelihood. To be precise, at each period we compare the individual models based on their predictive likelihoods over the past eight quarters and choose the best forecasting

model. The results from this model selection strategy is denoted as BMS.[6]

For each MCMC run, 20000 posterior draws are obtained after a burn-in period of length 5000. Geweke's diagnostic is used to check the convergence of the samplers. Table 1 reports the performance of the competing models for producing 1-quarter-ahead joint density forecasts, as well as the marginal density forecasts for the two components. These values are presented relative to those of the random walk benchmark: $\mathbf{y}_t = \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$, where $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Hence, any positive values indicate better forecast performance than the benchmark and vice versa.

A few broad conclusions can be drawn from the results.[7] For joint density forecasts, adding a moving average component improves the forecast performance of standard VARs for all lag lengths considered. For example, the difference between the log predictive likelihoods of VAR(3) and VARMA(3,1) is 2.8, which may be interpreted as a Bayes factor of about 16 in favor of the VARMA(3,1) model.

Table 1: Relative log predictive likelihoods for 1-quarter-ahead density forecasts (compared to the random walk model).

	Inflation	GDP	Joint
VAR(1)	73.7	24.6	99.9
VARMA(1,1)	86.7	23.1	109.2
VAR(2)	89.2	24.7	114.5
VAR(2)-SV	116.6	24.6	144.3
TVP-VAR(2)-SV	124.6	34.7	157.5
VARMA(2,1)	90.5	22.1	114.7
VARMA(2,1)-SV1	124.8	34.6	161.3
VARMA(2,1)-SV2	125.2	34.2	161.9
TVP-VARMA(2,1)-SV2	122.7	32.1	158.6
VAR(3)	86.0	22.3	111.0
VAR(3)-SV	115.4	23.5	142.6
TVP-VAR(3)-SV	122.2	34.4	161.1
VARMA(3,1)	89.0	21.7	113.8
VARMA(3,1)-SV1	121.7	33.6	159.0
VARMA(3,1)-SV2	122.5	32.6	159.3
TVP-VARMA(3,1)-SV2	119.6	28.4	152.7
BMS	119.8	35.9	149.0

In addition, adding stochastic volatility to the VARMAs substantially improves their forecast performance. This is in line with the large literature on inflation forecasting that shows the considerable benefits of allowing for stochastic volatility for both point and

[6]We thank an anonymous referee for this suggestion.

[7]The point forecast performance of the VARMAs is similar to their VAR counterparts. This might not be surprising as each pair of VARMA and VAR differs only in the autocovariance structure, which has a larger impact on the density forecasts than on the point forecasts.

density forecasts (see, e.g., Stock and Watson, 2007; Chan, Koop, Leon-Gonzalez, and Strachan, 2012; Clark and Doh, 2014). It is also interesting to note that VARMA(p,1)-SV1 and the more general VARMA(p,1)-SV2 perform very similarly, indicating that allowing for time-variation in $\mathbf{\Omega}_t$ is mostly sufficient. Overall, VARMA(2,1)-SV2 forecasts better than all other models.

Consistent with the results in D'Agostino, Gambetti, and Giannone (2013), allowing the VAR coefficients to be time-varying further improves the forecast performance of VAR(p)-SV. Interestingly, this does not hold for VARMAs: adding time-varying VAR coefficients slightly worsens the forecast performance of VARMA(p,1)-SV2. One possible explanation is that given the VMA coefficients are time-varying, ignoring this time-variation results in nonlinearities in the VAR coefficients. By explicitly modeling time-variation in the VMA coefficients, we can keep the VAR coefficients to be constant.

Next, to investigate the source of differences in forecast performance, we look at the log predictive likelihoods for each series. The results suggest that the gain in adding the moving average and stochastic volatility components comes mainly from forecasting inflation better, although adding stochastic volatility also improves forecasting GDP growth to some extent. Finally, the model selection strategy based on past forecast performance is highly competitive—it gives the best forecasts for GDP growth. Even when it is not the best performing model, its performance is comparable to the best.

Table 2 and Table 3 present the results for 2- and 3-quarter-ahead density forecasts, respectively. For longer horizons, the advantage of VARMAs over VARs is less substantial—it is perhaps not surprising as the VMA has only one lag. What remains to be important is allowing for stochastic volatility. In particular, both VAR(p)-SV and VARMA(p,1)-SV1 substantially outperform their counterparts without stochastic volatility. Among the two, the former models perform slightly better for 2-quarter-ahead forecasts, whereas the latter are better for 3-quarter-ahead forecasts.

Similar to 1-quarter-ahead results, allowing the VAR coefficients to be time-varying further improves the forecast performance of VAR(p)-SV, but this is not true for VARMAs. Again, the model selection strategy based on past forecast performance forecasts remarkably well. It often has the best forecasting performance.

Table 2: Relative log predictive likelihoods for 2-quarter-ahead density forecasts (compared to the random walk model).

	Inflation	GDP	Joint
VAR(1)	48.0	52.1	101.2
VARMA(1,1)	57.4	52.2	110.3
VAR(2)	63.8	53.1	117.7
VAR(2)-SV	84.2	54.7	142.3
TVP-VAR(2)-SV	89.6	63.8	151.7
VARMA(2,1)	65.8	51.0	118.4
VARMA(2,1)-SV1	83.3	54.6	142.0
VARMA(2,1)-SV2	83.1	55.2	141.7
TVP-VARMA(2,1)-SV2	80.6	55.2	137.3
VAR(3)	60.1	52.2	114.6
VAR(3)-SV	83.8	53.7	141.3
TVP-VAR(3)-SV	86.8	65.7	155.8
VARMA(3,1)	62.7	51.1	116.6
VARMA(3,1)-SV1	81.7	53.3	137.3
VARMA(3,1)-SV2	82.0	50.9	137.1
TVP-VARMA(3,1)-SV2	78.9	53.1	136.6
BMS	90.8	62.0	154.4

Table 3: Relative log predictive likelihoods for 3-quarter-ahead density forecasts (compared to the random walk model).

	Inflation	GDP	Joint
VAR(1)	43.1	75.0	117.9
VARMA(1,1)	42.7	75.1	118.3
VAR(2)	43.5	74.7	118.0
VAR(2)-SV	68.7	74.0	144.2
TVP-VAR(2)-SV	74.1	81.1	158.6
VARMA(2,1)	46.7	72.7	119.7
VARMA(2,1)-SV1	74.3	77.3	153.9
VARMA(2,1)-SV2	73.3	77.4	154.4
TVP-VARMA(2,1)-SV2	73.0	75.4	151.2
VAR(3)	41.6	73.5	115.0
VAR(3)-SV	69.6	72.5	144.1
TVP-VAR(3)-SV	70.6	82.0	154.2
VARMA(3,1)	49.6	70.5	120.7
VARMA(3,1)-SV1	74.1	72.0	146.7
VARMA(3,1)-SV2	73.6	70.0	145.4
TVP-VARMA(3,1)-SV2	72.9	68.4	144.2
BMS	71.8	80.8	158.9

To investigate the forecast performance of various competing models in more detail, we plot in Figure 1 the cumulative sums of log predictive likelihoods over the whole evaluation period (relative to the random walk model) for 1-quarter-ahead forecasts. It is clear from the figure that overall VARMAs—with or without stochastic volatility—consistently perform better than VARs.

The figure also reveals some interesting patterns over time. For example, during the Great Recession, the performance of both VARs and VARMAs with constant volatility deteriorates against the random walk model, whereas models with stochastic volatility perform substantially better. This again highlights the importance of allowing for stochastic volatility, especially during turbulent times.

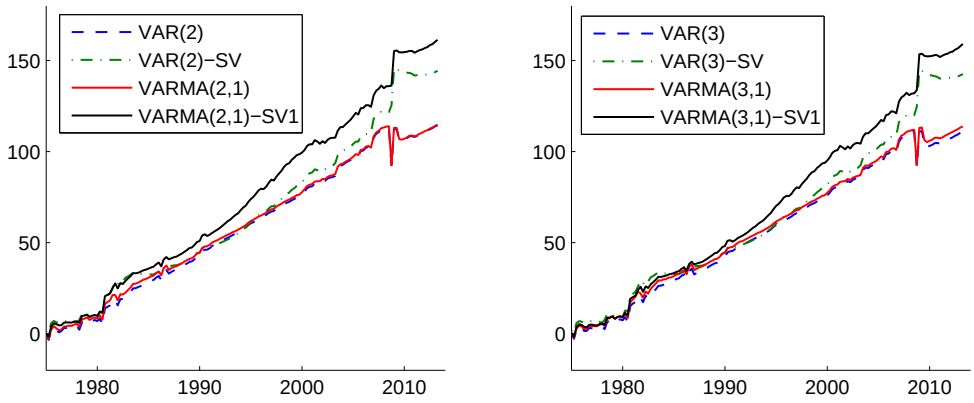

Figure 1: Cumulative sums of log predictive likelihoods for jointly forecasting inflation and GDP growth relative to the random walk model; one-quarter-ahead forecasts.

5 Concluding Remarks and Future Research

We build on a recently introduced latent factors representation of the VARMA model and derive some theoretical properties of the *expanded VARMA form* that justify its use in a Bayesian framework. On this foundation, we have developed a straightforward Gibbs sampler for the model and discussed how this algorithm can be extended to models with time-varying VMA coefficients and stochastic volatility. The proposed methodology was demonstrated using a density forecasting exercise, in which we also showed that VARMAs with stochastic volatility forecast better standard VARs with stochastic volatility.

The methodology developed in this article leads to a few lines of inquiry in the future. In Chan, Eisenstat, and Koop (2016), we develop algorithms that facilitate exact inference from echelon form VARMA specifications (with unknown Kronecker indices) and demonstrate that these can be readily used to estimate systems with as many as twelve equations. An interesting result we obtain is that moving average components and the canonical form gain in importance as the system size increases, even when compared to

parsimonious Bayesian VARs with shrinkage priors. However, in that work we do not consider extensions such as stochastic volatility, and this would be one point of interest for further research.

In addition, empirical investigation of alternative flexible specifications such as regime-switching VARMAs also seems beneficial. Moreover, it is worthwhile to further explore the relationship of the expanded form to dynamic factor models discussed in Section 2. Specifically, it can be shown that reducing the number of "factors" in this VARMA representation together with a particular set of restriction on the model parameters leads to exactly the FAVAR specification. It would therefore be of interest to further investigate how this relates to the recent work of Dufour and Stevanović (2013), as well as how alternative identification restrictions compare to the standard ones used in the literature. Lastly, we have focused on forecasting in this paper. Since flexible VARs are now routinely used for structural analysis, an interesting line of research is to perform similar analysis using VARMAs with time-varying parameters and stochastic volatility. We leave this for further research.

References

S. An and F. Schorfheide. Bayesian analysis of DSGE models. *Econometric Reviews*, 26 (2-4):113–172, 2007.

G. Athanasopoulos and F. Vahid. VARMA versus VAR for macroeconomic forecasting. *Journal of Business and Economic Statistics*, 26(2):237–252, 2008.

G. Athanasopoulos, D. S. Poskitt, and F. Vahid. Two canonical VARMA forms: Scalar component models vis-à-vis the echelon form. *Econometric Reviews*, 31(1):60–83, 2012.

M. Banbura, D. Giannone, and L. Reichlin. Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

M. A. G. Belmonte, G. Koop, and D. Korobilis. Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting*, 33(1):80–94, 2014.

B. Bernanke, J. Boivin, and P. S. Eliasz. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, 120(1):387–422, 2005.

F. Canova. Modelling and forecasting exchange rates with a Bayesian time-varying coefficient model. *Journal of Economic Dynamics and Control*, 17:233–261, 1993.

F. Canova. *Methods for Applied Macroeconomic Research*. Princeton University Press, New Jersey, 2007.

J. C. C. Chan. Moving average stochastic volatility models with application to inflation forecast. *Journal of Econometrics*, 176(2):162–172, 2013.

J. C. C. Chan and I. Jeliazkov. Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation*, 1:101–120, 2009.

J. C. C. Chan, G. Koop, R. Leon-Gonzalez, and R. Strachan. Time varying dimension models. *Journal of Business and Economic Statistics*, 30:358–367, 2012.

J. C. C. Chan, E. Eisenstat, and G. Koop. Large Bayesian VARMAs. *Journal of Econometrics*, 192(2):374–390, 2016.

T. E. Clark. Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business and Economic Statistics*, 29(3):327–341, 2011.

T. E. Clark and T. Doh. A Bayesian evaluation of alternative models of trend inflation. *International Journal of Forecasting*, 30(3):426–448, 2014.

T. Cogley and T. J. Sargent. Drifts and volatilities: monetary policies and outcomes in the post WWII US. *Review of Economic Dynamics*, 8(2):262 – 302, 2005.

T. F. Cooley and M. Dwyer. Business cycle analysis without much theory: A look at structural VARs. *Journal of Econometrics*, 83(12):57–88, 1998.

A. D'Agostino, L. Gambetti, and D. Giannone. Macroeconomic forecasting and structural change. *Journal of Applied Econometrics*, 28:82–101, 2013.

M. Del Negro and G. E. Primiceri. Time varying structural vector autoregressions and monetary policy: A corrigendum. *The Review of Economic Studies*, 2015. Forthcoming.

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

J.-M. Dufour and D. Pelletier. Practical methods for modelling weak VARMA processes: identification, estimation and specification with a macroeconomic application. Discussion Paper, McGill University, 2014.

J.-M. Dufour and D. Stevanović. Factor-augmented VARMA models with macroeconomic applications. *Journal of Business & Economic Statistics*, 31(4):491–506, 2013.

J. Durbin and S. J. Koopman. A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89:603–615, 2002.

E. Eisenstat, J. C. C. Chan, and R. W. Strachan. Model specification search for time-varying parameter VARs. *Econometric Reviews*, 35(8-10):1638–1665, 2016.

J. C. Engwerda, A. C. M. Ran, and A. L. Rijkeboer. Necessary and sufficient conditions for the existence of a positive definite solution of the matrix equation $X+A^*X^{-1}A = Q$. *Linear Algebra and its Applications*, 186:255–275, 1993.

J. Fernández-Villaverde, J. F. Rubio-Ramírez, T. J. Sargent, and M. W. Watson. ABCs (and Ds) of understanding VARs. *American Economic Review*, 97(3):1021–1026, 2007.

S. Fisk. A note on Weyl's inequality. *The American Mathematical Monthly*, 104(3): 257–258, 1997.

S. Frühwirth-Schnatter and H. Wagner. Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, 154:85–100, 2010.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

J. Geweke and G. Amisano. Hierarchical Markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics*, 26:1–29, 2011.

J. Ghosh and D. B. Dunson. Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320, 2009.

P. Gustafson. On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science*, 20(2):111–140, 2005.

J. D. Hamilton and G. Lin. Stock market volatility and the business cycle. *Journal of Applied Econometrics*, 11:573–593, 1996.

E. J. Hannan. The identification and parameterization of ARMAX and state space forms. *Econometrica*, 44(4):713–723, 1976.

N. J. Higham and H.-M. Kim. Numerical analysis of a quadratic matrix equation. *IMA Journal of Numerical Analysis*, 20:499–519, 2002.

K. Imai and D. A. van Dyk. A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, 124(2):311–334, 2005.

C. Kascha. A comparison of estimation methods for vector autoregressive moving-average models. *Econometric Reviews*, 31(3):297–324, 2012.

C. Kascha and C. Trenkler. Simple identification and specification of cointegrated VARMA models. *Journal of Applied Econometrics*, 2014. Forthcoming.

S. Kim, N. Shepherd, and S. Chib. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, 65(3):361–393, 1998.

G. Koop. Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, 2011. DOI: 10.1002/jae.1270.

G. Koop and D. Korobilis. Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, 3(4):267–358, 2010.

G. Koop and D. Korobilis. Large time-varying parameter VARs. *Journal of Econometrics*, 177(2):185–198, 2013.

G. Koop and S. M. Potter. Estimation and forecasting in models with multiple breaks. *Review of Economic Studies*, 74:763–789, 2007.

G. Koop, R. León-González, and R. W. Strachan. Efficient posterior simulation for cointegrated models with priors on the cointegration space. *Econometric Reviews*, 29 (2):224–242, 2010.

G. Koop, R. León-González, and R. W. Strachan. Bayesian model averaging in the instrumental variable regression model. *Journal of Econometrics*, 171(2):237 – 250, 2012.

D. Korobilis. Assessing the transmission of monetary policy shocks using time-varying parameter dynamic factor models. *Oxford Bulletin of Economics and Statistics*, 2012. Forthcoming.

D. Korobilis. VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics*, 28(2):204–230, 2013.

D. Korobilis. Data-based priors for vector autoregressions with drifting coefficients. *Available at SSRN 2392028*, 2014.

D. P. Kroese and J. C. C. Chan. *Statistical Modeling and Computation*. Springer, New York, 2014.

D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo Methods*. John Wiley & Sons, New York, 2011.

E. M. Leeper, T. B. Walker, and S.-C. S. Yang. Fiscal foresight: Analytics and econometrics. NBER Working Paper 14028, 2008.

H. Li and R. S. Tsay. A unified approach to identifying multivariate time series models. *Journal of the American Statistical Association*, 93(442):770–782, 1998.

M. Lippi and L. Reichlin. VAR analysis, nonfundamental representations, blaschke matrices. *Journal of Econometrics*, 63(1):307–325, 1994.

R. Litterman. Forecasting with Bayesian vector autoregressions — five years of experience. *Journal of Business and Economic Statistics*, 4:25–38, 1986.

J. S. Liu and Y. N. Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274, 1999.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, 2005.

H. Lütkepohl and H. Claessen. Analysis of cointegrated VARMA processes. *Journal of Econometrics*, 80:223–239, 1997.

H. Lütkepohl and D. S. Poskitt. Specification of echelon-form VARMA models. *Journal of Business & Economic Statistics*, 14(1):69–79, 1996.

X.-L. Meng and D.A. van Dyk. Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86(2):301–320, 1999.

K. Metaxoglou and A. Smith. Maximum likelihood estimation of VARMA models using a state-space EM algorithm. *Journal of Time Series Analysis*, 28(5):666–685, 2007.

R. Paap and H. van Dijk. Bayes estimates of Markov trends in possibly cointegrated series: An application to us consumption and income. *Journal of Business and Economic Statistics*, 21:547–563, 2003.

M. S. Peiris. On the study of some functions of multivariate ARMA processes. *Journal of Time Series Analysis*, 25(1):146–151, 1988.

D. J. Poirier. *Intermediate Statistics and Econometrics: A Comparative Approach*. MIT Press, Cambridge, 1995.

D. Poskitt and W. Yao. VAR modeling and business cycle analysis: A taxonomy of errors. *Journal of Business & Economic Statistics*, 2016.

D. S. Poskitt. Vector autoregressive moving average identification for macroeconomic modeling: A new methodology. *Journal of Econometrics*, 192(2):468–484, 2016.

G. E. Primiceri. Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies*, 72(3):821–852, 2005.

N.i Ravishanker and B. K. Ray. Bayesian analysis of vector ARMA models using Gibbs sampling. *Journal of Forecasting*, 16(3):177–194, 1997.

C. A. Sims. Macroeconomics and reality. *Econometrica*, 48:1–48, 1980.

J. H. Stock and M. W. Watson. Why has U.S. inflation become harder to forecast? *Journal of Money Credit and Banking*, 39:3–33, 2007.

S.-C. S. Yang. Quantifying tax effects under policy foresight. *Journal of Monetary Economics*, 52(8):1557 – 1568, 2005.

A Online Appendix

A.1 Solving the Matrix Quadratic Equation

Following Higham and Kim (2002), solutions to (12) are straightforward to obtain using *generalized Schur decompositions*. Specifically, define the $2n \times 2n$ matrices

$$\mathbf{F} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_n \\ -\widetilde{\mathbf{\Gamma}}_1' & \widetilde{\mathbf{\Gamma}}_0 \end{pmatrix}, \qquad \mathbf{G} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{\Gamma}}_1 \end{pmatrix},$$

and compute the decomposition $\mathbf{F} = \mathbf{QSZ}^*$, $\mathbf{G} = \mathbf{QTZ}^*$, where $\mathbf{Q}$ and $\mathbf{Z}$ are orthonormal, $\mathbf{S}$ and $\mathbf{T}$ are upper-triangular, and * denotes the conjugate transpose.[8] Next partition

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_{11} & \mathbf{Z}_{12} \\ \mathbf{Z}_{21} & \mathbf{Z}_{22} \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{0} & \mathbf{S}_{22} \end{pmatrix} \quad \mathbf{T} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{T}_{22} \end{pmatrix}.$$

Theorem 3 in Higham and Kim (2002) states that every solution to (12) has the form

$$\widehat{\mathbf{\Theta}}_1' = \mathbf{Z}_{21}\mathbf{Z}_{11}^{-1} = \mathbf{Q}_{11}\mathbf{S}_{11}\mathbf{T}_{11}^{-1}\mathbf{Q}_{11}^{-1}. \tag{21}$$

An important feature of the generalized Schur decomposition is that $\delta_i = s_i/t_i$ (where s_i and t_i are the i-th diagonal elements of $\mathbf{S}$ and $\mathbf{T}$, respectively) is a *generalized eigenvalue* of the pair $(\mathbf{F}, \mathbf{G})$, and if $1 \leq i \leq n$, it is also an eigenvalue of $\widehat{\mathbf{\Theta}}_1$. Since

$$\det(\mathbf{F} - \delta\mathbf{G}) = \det\left(\delta^2\widetilde{\mathbf{\Gamma}}_1 - \delta\widetilde{\mathbf{\Gamma}}_0 + \widetilde{\mathbf{\Gamma}}_1'\right) = \delta^2 \det\left(\left(\frac{1}{\delta}\right)^2\widetilde{\mathbf{\Gamma}}_1 - \frac{1}{\delta}\widetilde{\mathbf{\Gamma}}_0 + \widetilde{\mathbf{\Gamma}}_1'\right), \tag{22}$$

all generalized eigenvalues come in pairs $(\delta, 1/\delta)$, including the pairs $(0, \infty)$. Therefore, there will be exactly n generalized eigenvalues $|\delta_i| < 1$ and exactly n generalized eigenvalues $|\delta_i| > 1$.

Note that we are free to rotate the rows and columns of $\mathbf{Q}$, $\mathbf{Z}$, $\mathbf{S}$, $\mathbf{T}$ in forming the solution for $\widehat{\mathbf{\Theta}}_1$, and so we may choose any rotation that retains a desired subset of n generalized eigenvalues of $(\mathbf{F}, \mathbf{G})$ to be the eigenvalues of $\widehat{\mathbf{\Theta}}_1$. For example, to obtain the fundamental representation, we would choose the rotation that yields $\mathbf{S}_{11}\mathbf{T}_{11}^{-1}$ with all diagonal elements $|s_i/t_i| < 1$.

A.2 Proofs of Theorems

Proof of Theorem 1. Let $\mathbf{u}_t = \mathbf{\Theta}(L)\boldsymbol{\varepsilon}_t$ be a VMA(q) process with $\boldsymbol{\varepsilon}$ being weak white noise and no roots of $\det\mathbf{\Theta}(L)$ lying on the unit circle. Then $\{\mathbf{u}_t\}$ is a *purely non-deterministic* process in a Hilbert space spanned by $\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_{t-1}, \ldots$, i.e., $\mathbf{u}_t \in \mathcal{L}^2(\boldsymbol{\varepsilon}; t)$, and $\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_{t-1}, \ldots$ is a purely non-deterministic process, i.e., $\lim_{t\to-\infty} \mathcal{L}^2(\boldsymbol{\varepsilon}; t) = \{\mathbf{0}\}$.

[8] Many modern statistical packages have built-in functions for computing generalized Schur decompositions—in our empirical work, we use the command `qz` in MATLAB.

We decompose $\mathcal{L}^2(\boldsymbol{\varepsilon};t)$ by an orthogonal projection such that $\mathcal{L}^2(\boldsymbol{\varepsilon};t) = \mathcal{L}^2(\boldsymbol{\nu};t) \bigoplus \mathcal{L}^2(\boldsymbol{\zeta};t)$, and for every $\mathbf{u}_t \in \mathcal{L}^2(\boldsymbol{\varepsilon};t)$ we can write

$$\mathbf{u}_t = \boldsymbol{\Upsilon}(L)\boldsymbol{\nu}_t + \boldsymbol{\Xi}(L)\boldsymbol{\zeta}_t, \tag{23}$$

where $\boldsymbol{\nu}_t$ and $\boldsymbol{\zeta}_t$ are $n \times 1$, $\mathrm{E}(\boldsymbol{\nu}_t) = \mathrm{E}(\boldsymbol{\zeta}_t) = \mathbf{0}$, $\mathrm{E}(\boldsymbol{\nu}_t\boldsymbol{\nu}_t') = \mathrm{E}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t') = \mathbf{I}_n$, $\mathrm{E}(\boldsymbol{\nu}_t\boldsymbol{\nu}_{t-s}') = \mathrm{E}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_{t-s}') = \mathbf{0}$ for all $s \geq 1$, and $\mathrm{E}(\boldsymbol{\nu}_t\boldsymbol{\zeta}_{t-s}') = \mathbf{0}$ for all $s \geq 0$. $\boldsymbol{\Upsilon}(L)$ and $\boldsymbol{\Xi}(L)$ are some finite-order $n \times n$ polynomial matrices in the lag operator satisfying $\operatorname{rank}\boldsymbol{\Upsilon}(L) + \operatorname{rank}\boldsymbol{\Xi}(L) > n$. The latter follows from the fact that any univariate MA with no roots on the unit circle can be obviously decomposed this way, and any $n \times n$ polynomial matrix $\boldsymbol{\Theta}(L)$ can be written as the product $\mathcal{E}(L)\boldsymbol{\Theta}^{\dagger}(L)$, where $\mathcal{E}(L)$ is a product of *elementary matrices* such that $\det\mathcal{E}(L)$ is constant (i.e., $\mathcal{E}(L)$ is unimodular) and $\boldsymbol{\Theta}^{\dagger}(L)$ is upper-triangular.

Without loss of generality, assume $\boldsymbol{\Xi}_0 \equiv \boldsymbol{\Xi}(0)$ is invertible. In particular, if neither $\boldsymbol{\Xi}_0$ nor $\boldsymbol{\Upsilon}_0$ are invertible, we can always construct a $2n \times 2n$ Blaschke matrix $\mathbf{C}(L)$ (with $\mathbf{C}(L^{-1})'\mathbf{C}(L) = \mathbf{I}_{2n}$) such that

$$\begin{aligned}
\begin{pmatrix} \widetilde{\boldsymbol{\nu}}_t \\ \widetilde{\boldsymbol{\zeta}}_t \end{pmatrix} &= \begin{pmatrix} \mathbf{C}_{11}(L) & \mathbf{C}_{12}(L) \\ \mathbf{C}_{21}(L) & \mathbf{C}_{22}(L) \end{pmatrix} \begin{pmatrix} \boldsymbol{\nu}_t \\ \boldsymbol{\zeta}_t \end{pmatrix}, \\
\widetilde{\boldsymbol{\Upsilon}}(L) &= \boldsymbol{\Upsilon}(L)\mathbf{C}_{11}(L^{-1})' + \boldsymbol{\Xi}(L)\mathbf{C}_{12}(L^{-1})', \\
\widetilde{\boldsymbol{\Xi}}(L) &= \boldsymbol{\Upsilon}(L)\mathbf{C}_{21}(L^{-1})' + \boldsymbol{\Xi}(L)\mathbf{C}_{22}(L^{-1})',
\end{aligned}$$

yields an equivalent representation

$$\mathbf{u}_t = \widetilde{\boldsymbol{\Upsilon}}(L)\widetilde{\boldsymbol{\nu}}_t + \widetilde{\boldsymbol{\Xi}}(L)\widetilde{\boldsymbol{\zeta}}_t, \tag{24}$$

but with $\widetilde{\boldsymbol{\Xi}}_0 \equiv \widetilde{\boldsymbol{\Xi}}(0)$ invertible (and $\widetilde{\boldsymbol{\Upsilon}}(L) \neq 0$).[9]

Proceeding under the assumption that $\boldsymbol{\Xi}_0$ is invertible, further decompose $\boldsymbol{\Xi}_0\boldsymbol{\zeta}_t = \mathbf{K}\boldsymbol{\xi}_t + \boldsymbol{\eta}_t$, where $\boldsymbol{\xi}_t$ and $\boldsymbol{\eta}_t$ are $n \times 1$, $\mathrm{E}(\boldsymbol{\xi}_t) = \mathrm{E}(\boldsymbol{\eta}_t) = \mathbf{0}$, $\mathrm{E}(\boldsymbol{\xi}_t\boldsymbol{\xi}_t') = \mathbf{I}_n$, $\mathrm{E}(\boldsymbol{\eta}_t\boldsymbol{\eta}_t') = \boldsymbol{\Lambda}$, $\mathrm{E}(\boldsymbol{\xi}_t\boldsymbol{\xi}_{t-s}') = \mathrm{E}(\boldsymbol{\eta}_t\boldsymbol{\eta}_{t-s}') = \mathbf{0}$ for all $s \geq 1$, and $\mathrm{E}(\boldsymbol{\xi}_t\boldsymbol{\eta}_{t-s}') = \mathbf{0}$ for all $s \geq 0$. $\boldsymbol{\Lambda}$ is diagonal with elements $\lambda_i^2 > 0$.

Let $\widehat{\mathbf{u}}_t = \sum_{j=0}^{q_\nu} \boldsymbol{\Upsilon}_j\boldsymbol{\nu}_{t-j} + \sum_{j=1}^{q_\zeta} \boldsymbol{\Xi}_j\boldsymbol{\zeta}_{t-j} + \mathbf{K}\boldsymbol{\xi}_t$, such that $\mathbf{u}_t = \widehat{\mathbf{u}}_t + \boldsymbol{\eta}_t$. Observe that $\{\widehat{\mathbf{u}}_t\}$ is a purely non-deterministic, stationary process with the properties $\mathrm{E}(\widehat{\mathbf{u}}_t\widehat{\mathbf{u}}_{t-s}') = 0$ for all $s > \max\{q_\nu, q_\zeta\}$ and $\mathrm{E}(\widehat{\mathbf{u}}_t\boldsymbol{\eta}_{t-s}') = 0$ for all $s \geq 0$. Therefore, employing the Wold decomposition yields

$$\widehat{\mathbf{u}}_t = \sum_{j=0}^{\widehat{q}} \boldsymbol{\Pi}_j\mathbf{g}_t, \tag{25}$$

where $\mathbf{g}_t$ is $n \times 1$, $\mathrm{E}(\mathbf{g}_t) = 0$, $\mathrm{E}(\mathbf{g}_t\mathbf{g}_t') = \boldsymbol{\Psi}$, $\mathrm{E}(\mathbf{g}_t\mathbf{g}_{t-s}') = 0$ for all $s \geq 1$ and $\mathrm{E}(\mathbf{g}_t\boldsymbol{\eta}_{t-s}) = 0$ for all $s \geq 0$. $\boldsymbol{\Pi}_0 = \mathbf{I}_n$ and $\boldsymbol{\Psi}$ is positive semi-definite (p.s.d.).

[9]Recall that applying a Blaschke transformation to any orthonormal white noise vector $\mathbf{w}_t$, via the transformation $\widetilde{\mathbf{w}}_t = \mathbf{C}(L)\mathbf{w}_t$, results in an orthonormal white nose vector $\widetilde{\mathbf{w}}_t$ (see Lippi and Reichlin, 1994, for further details).

Take the LDU decomposition of $\boldsymbol{\Psi} = \boldsymbol{\Phi}_0\boldsymbol{\Omega}\boldsymbol{\Phi}_0'$ and set $\boldsymbol{\Phi}_j = \boldsymbol{\Pi}_j\boldsymbol{\Phi}_0$ for $j = 0, \ldots, \widehat{q}$, $\mathbf{f}_t = \boldsymbol{\Phi}_0^{-1}\mathbf{g}_t$. Since

$$\mathrm{E}(\mathbf{u}_t\mathbf{u}_{t-s}') = \mathrm{E}((\widehat{\mathbf{u}}_t + \boldsymbol{\eta}_t)(\widehat{\mathbf{u}}_{t-s} + \boldsymbol{\eta}_{t-s})')$$

for all $s \geq 0$ is a necessary condition, it must be that $\widehat{q} = q$ and we obtain the representation (4), which satisfies (6). □

Proof of Theorem 2. Similar to the VMA(1) representation (2) of a VMA(q), define the companion representation of the expanded form by

$$\underbrace{\begin{pmatrix} \mathbf{u}_t \\ \mathbf{u}_{t-1} \\ \vdots \\ \mathbf{u}_{t-q+1} \end{pmatrix}}_{\widetilde{\mathbf{u}}_\tau} = \underbrace{\begin{pmatrix} \boldsymbol{\Phi}_0 & \boldsymbol{\Phi}_1 & \cdots & \boldsymbol{\Phi}_{q-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \boldsymbol{\Phi}_1 \\ & & & \boldsymbol{\Phi}_0 \end{pmatrix}}_{\widetilde{\boldsymbol{\Phi}}_0} \underbrace{\begin{pmatrix} \mathbf{f}_t \\ \mathbf{f}_{t-1} \\ \vdots \\ \mathbf{f}_{t-q+1} \end{pmatrix}}_{\widetilde{\mathbf{f}}_\tau}$$
$$+ \underbrace{\begin{pmatrix} \boldsymbol{\Phi}_q & & & \\ \boldsymbol{\Phi}_{q-1} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \boldsymbol{\Phi}_1 & \cdots & \boldsymbol{\Phi}_{q-1} & \boldsymbol{\Phi}_q \end{pmatrix}}_{\widetilde{\boldsymbol{\Phi}}_1} \underbrace{\begin{pmatrix} \mathbf{f}_{t-q} \\ \mathbf{f}_{t-q-1} \\ \vdots \\ \mathbf{f}_{t-2q+1} \end{pmatrix}}_{\widetilde{\mathbf{f}}_{\tau-1}} + \underbrace{\begin{pmatrix} \boldsymbol{\eta}_t \\ \boldsymbol{\eta}_{t-1} \\ \vdots \\ \boldsymbol{\eta}_{t-q+1} \end{pmatrix}}_{\widetilde{\boldsymbol{\eta}}_\tau} \tag{26}$$

with

$$\begin{pmatrix} \widetilde{\mathbf{f}}_\tau \\ \widetilde{\boldsymbol{\eta}}_\tau \end{pmatrix} \sim \mathcal{WN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{I}_q \otimes \boldsymbol{\Omega} & 0 \\ 0 & \mathbf{I}_q \otimes \boldsymbol{\Lambda} \end{pmatrix} \right). \tag{27}$$

Accordingly, the mapping in (6) can be written as:

$$\widetilde{\boldsymbol{\Theta}}_0(\mathbf{I}_q \otimes \boldsymbol{\Sigma})\widetilde{\boldsymbol{\Theta}}_0' + \widetilde{\boldsymbol{\Theta}}_1(\mathbf{I}_q \otimes \boldsymbol{\Sigma})\widetilde{\boldsymbol{\Theta}}_1' = \widetilde{\boldsymbol{\Phi}}_0(\mathbf{I}_q \otimes \boldsymbol{\Omega})\widetilde{\boldsymbol{\Phi}}_0' + \widetilde{\boldsymbol{\Phi}}_1(\mathbf{I}_q \otimes \boldsymbol{\Omega})\widetilde{\boldsymbol{\Phi}}_1' + \widetilde{\boldsymbol{\Lambda}}, \tag{28}$$

$$\widetilde{\boldsymbol{\Theta}}_1(\mathbf{I}_q \otimes \boldsymbol{\Sigma})\widetilde{\boldsymbol{\Theta}}_0' = \widetilde{\boldsymbol{\Phi}}_1(\mathbf{I}_q \otimes \boldsymbol{\Omega})\widetilde{\boldsymbol{\Phi}}_0'. \tag{29}$$

Let $\boldsymbol{\Psi} = \widetilde{\boldsymbol{\Phi}}_0(\mathbf{I}_q \otimes \boldsymbol{\Omega})\widetilde{\boldsymbol{\Phi}}_0'$. Then,

$$\boldsymbol{\Phi}_0\boldsymbol{\Omega}\boldsymbol{\Phi}_0' = \begin{pmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}_n \end{pmatrix} \boldsymbol{\Psi} \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{I}_n \end{pmatrix}, \tag{30}$$

$$\begin{pmatrix} \boldsymbol{\Phi}_q \\ \vdots \\ \boldsymbol{\Phi}_1 \end{pmatrix} = \widetilde{\boldsymbol{\Theta}}_1(\mathbf{I}_q \otimes \boldsymbol{\Sigma})\widetilde{\boldsymbol{\Theta}}_0'\boldsymbol{\Psi}^{-1} \begin{pmatrix} \boldsymbol{\Phi}_0^{-1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}. \tag{31}$$

Since $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Omega}$ are obtained uniquely from (30) via the LDU decomposition, we see that there is a unique mapping from $\boldsymbol{\Psi}$ to $\boldsymbol{\Phi}_0, \ldots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}$. In other words, all information

about $\boldsymbol{\Phi}_0, \ldots, \boldsymbol{\Phi}_q, \boldsymbol{\Omega}$ is contained in $\boldsymbol{\Psi}$, so it suffices to analyze the mapping between $\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}, \boldsymbol{\Lambda}$.

To this end, substituting (29) into (28) leads to

$$\widetilde{\boldsymbol{\Gamma}}_1 \boldsymbol{\Psi}^{-1} \widetilde{\boldsymbol{\Gamma}}_1' + \boldsymbol{\Psi} = \widetilde{\boldsymbol{\Gamma}}_0 - \mathbf{I}_q \otimes \boldsymbol{\Lambda}, \tag{32}$$

where $\widetilde{\boldsymbol{\Gamma}}_0, \widetilde{\boldsymbol{\Gamma}}_1$ are computed from $\widetilde{\boldsymbol{\Theta}}_0, \widetilde{\boldsymbol{\Theta}}_1, \boldsymbol{\Sigma}$ as in (3). By definition of $\mathbb{P}$, the set of permissible expanded form parameters can be cast as all real positive definite $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$ that solve (32). More specifically, $\mathbb{M}_{\boldsymbol{\lambda}}$ corresponds to the set of all real positive definite solutions $\boldsymbol{\Psi}$ for a given $\boldsymbol{\lambda}$, and $\mathbb{L}$ is the set of all $\boldsymbol{\lambda}$ for which a positive definite solution $\boldsymbol{\Psi}$ to (32) exists.

Engwerda, Ran, and Rijkeboer (1993) demonstrate that a necessary condition for the existence of a positive definite solution is $\widetilde{\boldsymbol{\Gamma}}_0 + \rho \widetilde{\boldsymbol{\Gamma}}_1 + \rho^{-1} \widetilde{\boldsymbol{\Gamma}}_1' - \mathbf{I}_q \otimes \boldsymbol{\Lambda}$ must be positive semi-definite for all $|\rho| = 1$, which is equivalent to requiring $\mathbf{H}_\rho - \mathbf{I}_q \otimes \boldsymbol{\Lambda}$ to be p.s.d. for all $|\rho| = 1$. Part 1 of the theorem then follows directly from applying Theorem 1 of Fisk (1997) to the sum $-\mathbf{H}_\rho + \mathbf{I}_q \otimes \boldsymbol{\Lambda}$. Moreover, Proposition 8.2 of Engwerda, Ran, and Rijkeboer (1993) states that whenever a solution exists, there are at most a finite number of real-valued $\boldsymbol{\Psi}$ that solve (32), which proves part 2 of the theorem. Finally, part 1 and part 2 together imply that the set $\mathbb{P}$ is bounded.

To prove the corollary, note that if z^* is a root of $\boldsymbol{\Theta}(z)$ on the unit circle, then z^* and $\bar{z}^*$ are eigenvalues of $\widetilde{\boldsymbol{\Theta}}_1 \widetilde{\boldsymbol{\Theta}}_0^{-1}$. Hence, for $\rho = -\bar{z}^*$, the matrix $\mathbf{H}_\rho$ is singular, and there exists a $qn \times 1$ vector $\boldsymbol{\pi}$—e.g., the eigenvector of $\mathbf{H}_\rho$ associated with the 0 eigenvalue—such that $\boldsymbol{\pi}' \mathbf{H}_\rho \boldsymbol{\pi} = 0$ and the necessary condition for the existence of a positive definite solution to (32) reduces to

$$-\boldsymbol{\pi}' (\mathbf{I}_q \otimes \boldsymbol{\Lambda}) \boldsymbol{\pi} \geq 0,$$

which has the form

$$\sum_{i=1} \lambda_i^2 \sum_{j=1}^{q} \pi_{i,j}^2 = 0,$$

where $\pi_{i,j}$ is an element of $\boldsymbol{\pi}$. Since at least one $\pi_{i,j} \neq 0$, there must be at least one $\lambda_i^2 = 0$. If $\boldsymbol{\pi}$ does not contain any 0 elements, then $\lambda_i^2 = 0$ for all $i = 1, \ldots, n$. □

A.3 Details of Sampling Algorithm

In this appendix we discuss the details of Step 2 to Step 4 in the Gibbs sampler for the VARMA(p, q) model with time-varying parameters and stochastic volatilities in the expanded form given by (17).

To sample $(\boldsymbol{\phi} \,|\, \mathbf{y}, \boldsymbol{\beta}, \mathbf{f}, \mathbf{h}, \boldsymbol{\psi}_\phi^2, \boldsymbol{\psi}_h^2, \boldsymbol{\Lambda})$ in Step 2, note that the innovation $\boldsymbol{\eta}_t$ has a diagonal covariance matrix $\boldsymbol{\Lambda}$. Hence, we can estimate $\boldsymbol{\Phi}_{0,t}, \ldots, \boldsymbol{\Phi}_{q,t}$ equation by equation. To

that end, define $\mathbf{y}_t^* = \mathbf{y}_t - \boldsymbol{\mu} - \mathbf{A}_1\mathbf{y}_{t-1} - \cdots - \mathbf{A}_p\mathbf{y}_{t-p}$, and let $y_{i,t}^*$ denote the i-th element of $\mathbf{y}_t^*$.

As discussed in Section 2, we need to impose linear restrictions on the elements of $\boldsymbol{\Phi}_{0,t}$—in particular $\boldsymbol{\Phi}_{0,t}$ is a lower triangular matrix with ones on the main diagonal—and it will often be of interest to further impose exclusion restrictions on $\mathbf{A}_1, \ldots, \mathbf{A}_p$ (alternatively, $\mathbf{B}_0, \ldots, \mathbf{B}_p$ for the echelon form) and $\boldsymbol{\Phi}_{1,t}, \ldots, \boldsymbol{\Phi}_{q,t}$ as well. To economize on space, we explicitly discuss implementing such restrictions on $\boldsymbol{\Phi}_{0,t}, \ldots, \boldsymbol{\Phi}_{q,t}$; sampling $\mathbf{A}_1, \ldots, \mathbf{A}_p$ subject to similar restrictions follows analogously. Let $\boldsymbol{\phi}_{j,t,i}^\star$ denote the i-th row of $\boldsymbol{\Phi}_{j,t}$ and let $\boldsymbol{\phi}_{j,t,i}$ be the free elements in $\boldsymbol{\phi}_{j,t,i}^\star$, such that

$$\begin{aligned}\boldsymbol{\phi}_{0,t,i}^\star &= \mathbf{R}_{0,i}\boldsymbol{\phi}_{0,t,i} + \boldsymbol{\iota}_i, \\ \boldsymbol{\phi}_{j,t,i}^\star &= \mathbf{R}_{j,i}\boldsymbol{\phi}_{j,t,i}, \quad \text{for } j \geqslant 1,\end{aligned}$$

where $\boldsymbol{\iota}_i$ is an $n \times 1$ vector with the i-th element $\boldsymbol{\iota}_{ii} = 1$ and all others set to zero. $\mathbf{R}_{j,i}$ in this context is a pre-determined selection matrix of appropriate dimensions. For example, our assumption that $\boldsymbol{\Phi}_{0,t}$ is a lower triangular matrix with ones on the main diagonal corresponds to $\mathbf{R}_{0,i} = (\mathbf{I}_{i-1}, \mathbf{0})'$ for $i > 1$. For $i = 1$, $\boldsymbol{\phi}_{0,1} = \boldsymbol{\iota}_1$ and there are no free elements. If there are no restrictions imposed on $\boldsymbol{\Phi}_{1,t}, \ldots, \boldsymbol{\Phi}_{q,t}$, then $\mathbf{R}_{1,i} = \cdots = \mathbf{R}_{q,i} = \mathbf{I}_{nq}$. Using this formulation, define $\mathbf{R}_i = \text{diag}(\mathbf{R}_{0,1}, \mathbf{R}_{1,1}, \ldots, \mathbf{R}_{q,i})$ and $\boldsymbol{\phi}_{i,t} = (\boldsymbol{\phi}_{0,t,i}', \boldsymbol{\phi}_{1,t,i}' \ldots, \boldsymbol{\phi}_{q,t,i}')'$, such that

$$\boldsymbol{\phi}_{i,t}^\star = \mathbf{R}_i\boldsymbol{\phi}_{i,t} + \begin{pmatrix}\boldsymbol{\iota}_i \\ \mathbf{0}\end{pmatrix}$$

forms the i-th row of $(\boldsymbol{\Phi}_0, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_q)$.

Then, the i-th equation in (17) can be written as

$$y_{i,t}^* = f_{i,t} + \widetilde{\mathbf{w}}_t\mathbf{R}_i\boldsymbol{\phi}_{i,t} + \eta_{i,t}, \tag{33}$$

where $\widetilde{\mathbf{w}}_t = (\mathbf{f}_t', \mathbf{f}_{t-1}', \ldots, \mathbf{f}_{t-q}')$, $f_{i,t}$ is the i-th element of $\mathbf{f}_t$, and $\eta_{i,t}$ is the i-th element of $\boldsymbol{\eta}_t$. Hence, (18) and (33) define a linear Gaussian state space model for $\boldsymbol{\phi}_{i,t}$, and standard algorithms can be applied. Here we use the algorithm in Chan and Jeliazkov (2009) to jointly sample $\boldsymbol{\phi}_{i,1}, \ldots, \boldsymbol{\phi}_{i,T}$.

To implement Step 3, one can directly apply the algorithm in Del Negro and Primiceri (2015) that is based on the auxiliary mixture sampler of Kim, Shepherd, and Chib (1998) to draw the log-volatilities.

Lastly, Step 4 involves drawing from $p(\boldsymbol{\Lambda} \,|\, \mathbf{y}, \boldsymbol{\beta}, \mathbf{f}, \boldsymbol{\phi})$, $p(\boldsymbol{\psi}_h^2 \,|\, \mathbf{h})$ and $p(\boldsymbol{\psi}_\phi^2 \,|\, \boldsymbol{\phi})$. The first density is a product of inverse-gamma densities, and can therefore be sampled from using standard methods:

$$(\lambda_i^2 \,|\, \mathbf{y}, \boldsymbol{\beta}, \mathbf{f}, \boldsymbol{\phi}) \sim \mathcal{IG}\left(\nu_{\lambda,0} + T/2, S_{\lambda,0} + \sum_{t=1}^{T}(y_{i,t}^* - f_{i,t} - \widetilde{\mathbf{w}}_t\mathbf{R}_i\boldsymbol{\phi}_{i,t})^2/2\right).$$

The last two densities are nonstandard, but we can sample from them using a Metropolis-Hastings step with a tailored proposal density. For example, to sample $\psi_{h,i}^2$, we first obtain

a candidate draw $\psi^{2*}_{h,i}$ from the proposal $\mathcal{IG}((T-3)/2, \sum_{t=2}^{T}(h_{i,t}-h_{i,t-1})^2/2)$. Given the current draw $\psi^{2}_{h,i}$, the candidate $\psi^{2*}_{h,i}$ is accepted with probability

$$\min\left\{1, \frac{(\psi^{2*}_{h,i})^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2S_{h,0}}\psi^{2*}_{h,i}}}{(\psi^{2}_{h,i})^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2S_{h,0}}\psi^{2}_{h,i}}}\right\};$$

otherwise we keep the current draw $\psi^{2}_{h,i}$. We sample $\psi^{2}_{\phi,i,j}$ similarly.