AARHUS
UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

CREATES
Center for Research in Econometric Analysis of Time Series

Noncausal Bayesian Vector Autoregression

Markku Lanne and Jani Luoto

CREATES Research Paper 2014-7

Department of Economics and Business
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Noncausal Bayesian Vector Autoregression[†]

Markku Lanne
University of Helsinki and CREATES

Jani Luoto*
University of Helsinki

March 2014

Abstract

We propose a Bayesian inferential procedure for the noncausal vector autoregressive (VAR) model that is capable of capturing nonlinearities and incorporating effects of missing variables. In particular, we devise a fast and reliable posterior simulator that yields the predictive distribution as a by-product. We apply the methods to postwar quarterly U.S. inflation and GDP growth series. The noncausal VAR model turns out to be superior in terms of both in-sample fit and out-of-sample forecasting performance over its conventional causal counterpart. In addition, we find GDP growth to have predictive power for the future distribution of inflation over and above the own history of inflation, but not vice versa. This may be interpreted as evidence against the new Keynesian model that implies Granger causality from inflation to GDP growth, provided GDP growth is a reasonable proxy of the marginal cost.

Keywords: Noncausal time series, non-Gaussian time series, Bayesian analysis, New Keynesian model.
JEL classification: C11, C32, E31.

[†] Financial support from the Academy of Finland is gratefully acknowledged. The first author also acknowledges financial support from CREATES (DNRF78) funded by the Danish National Research Foundation, while the second author is grateful for financial support from the Yrjö Jahnsson Foundation and the OP-Pohjola Group Research Foundation.

* Corresponding author. Department of Political and Economic Studies, University of Helsinki, P.O.Box 17 (Arkadiankatu 7), FIN–00014 University of Helsinki, Finland, e-mail: jani.luoto@helsinki.fi

1 Introduction

While the vast majority of empirical analysis of multivariate time series in macroeconomics and finance is based on the linear vector autoregressive (VAR) model, there has been an increasing interest in nonlinear multivariate time series models in the last few decades, especially followed by the burgeoning literature on theoretical nonlinear macroeconomic models. One such model is the noncausal VAR model recently put forth by Davis and Song (2010), and Lanne and Saikkonen (2013). While these two specifications differ somewhat from each other, they are both characterized by the defining feature of any noncausal process of explicit dependence on the future such that the current value has no linear representation in terms of current and past errors. This tends to complicate interpretation as the errors of the noncausal VAR model become predictable from past observations, and hence, cannot be thought of as shocks in any economic sense. On the other hand, as pointed out by Lanne and Saikkonen (2013), the model has a nonlinear causal representation, and there is mounting evidence of the versatility of nonlinearity that noncausal models are capable of generating (see, e.g., Gourieroux and Zakoïan (2013), and Lof (2012)). Although little is known of the form of nonlinearity afforded by the noncausal VAR process, it can be seen as a convenient shorthand way of writing a complicated nonlinear model. In addtition to capturing nonlinearities, the noncausal VAR model is capable of incorporating effects of missing variables as suggested by Lanne and Saikkonen (2013), and it may, therefore, be useful in many macroeconomic and financial applications, where assessing the adequacy of the included set of variables tends to be problematic.

In this paper, we consider Bayesian analysis, including estimation and forecasting of the noncausal VAR model (see DelNegro and Schorfheide (2011) and Karlsson (2012) for recent reviews of the causal Bayesian VAR models). Our approach is an extension of Lanne, Luoma, and Luoto (2012), who proposed corresponding methods for the univariate noncausal autoregressive (AR) model. In particular, we show how the posterior density of the noncausal (and hence nonlinear) VAR model can be manipulated to facilitate estimation by a straightforward extension of the commonly employed Gibbs sampling algorithm of Kadiyala and Karlsson (1997). The resulting sampler also conveniently yields the posterior predictive distribution as a by-product. Forecasting in the noncausal VAR model has previously been considered by Nyberg and Saikkonen (2013), but their frequentist approach requires considerably more complicated techniques than ours.

We apply the noncausal VAR model to quarterly U.S. inflation and GDP growth

series (from 1955:1 to 2013:2), where clear evidence in favor of noncausality is detected. The noncausal VAR model also turns out to be superior in point and density forecasting. Finally, we devise a Bayesian procedure to test for Granger noncausality in distribution (see Droumaguet and Woźniak (2012)), and apply it to these two variables. According to the new Keynesian model, inflation should Granger cause GDP growth if the latter is a reasonable proxy of the marginal cost. In line with much of the previous literature, we find no evidence of Granger causality from inflation to GDP growth, which can be interpreted as evidence against the new Keynesian model. Interestingly, however, there is strong evidence of Granger causality from GDP growth to inflation, which has typically not been detected in analyses based on the linear causal VAR model. The latter finding points to the nonlinear nature of Granger causality in this case.

The plan of the rest of the paper is as follows. In Section 2, we review the noncausal VAR model of Lanne and Saikkonen (2013) and discuss its interpretation. In Section 3, we introduce the Bayesian estimation procedure, while in Section 4 it is extended to produce forecasts. The empirical application to U.S. inflation and GDP growth is presented in Section 5. Finally, Section 6 concludes.

2 Model

The n-dimensional VAR(r,s) process y_t $(t=0,\pm1,\pm2,...)$ proposed by Lanne and Saikkonen (2013) is generated by

$$\Pi\left(B\right)\Phi\left(B^{-1}\right)y_t=\epsilon_t, \tag{1}$$

where $\Pi\left(B\right)=I_n-\Pi_1B-\cdots-\Pi_rB^r$ $(n\times n)$ and $\Phi\left(B^{-1}\right)=I_n-\Phi_1B^{-1}-\cdots-\Phi_sB^{-s}$ $(n\times n)$ are matrix polynomials in the backward shift operator B, and ϵ_t $(n\times1)$ is a sequence of independent, identically distributed (continuous) random vectors with zero mean and finite positive definite covariance matrix. If $\Phi_j\neq0$ for some $j\in\{1,...,s\}$, equation (1) defines a noncausal vector autoregression referred to as purely noncausal when $\Pi_1=\cdots=\Pi_r=0$. The corresponding conventional causal model is obtained when $\Phi_1=\cdots=\Phi_s=0$, and in keeping with the conventional notation in the literature, we sometimes use the abbreviation VAR(r) in this case.

Stationarity of the process is guaranteed by the assumption that the matrix polynomials $\Pi\left(z\right)$ and $\Phi\left(z\right)$ $(z\in\mathbb{C})$ have their zeros outside the unit disc, i.e.,

$$\det\Pi\left(z\right)\neq0,\quad|z|\leq1,\quad\text{and}\quad\det\Phi\left(z\right)\neq0,\quad|z|\leq1. \tag{2}$$

Specifically, the process

$$u_t = \Phi\left(B^{-1}\right) y_t$$

is then stationary, and, as pointed out by Lanne and Saikkonen (2013), there exists a $\delta_1 > 0$ such that $\Pi(z)^{-1}$ has a well defined power series representation $\Pi(z)^{-1} = \sum_{j=0}^{\infty} M_j z^j = M(z)$ for $|z| < 1 + \delta_1$, indicating that the process u_t has the causal moving average representation

$$u_t = M(B)\,\epsilon_t = \sum_{j=0}^{\infty} M_j \epsilon_{t-j}. \tag{3}$$

Notice that $M_0 = I_n$ and that (the elements of) the coefficient matrices M_j decay to zero at a geometric rate as $j \to \infty$ (cf. Lemma 3 in Kohn (1979)). When convenient, $M_j = 0$, $j < 0$, will be assumed.

In the same vein, due to the latter condition in (2), the process $w_t = |\Pi(B)|\, y_t$ has the following representation

$$w_t = \sum_{j=-(n-1)r}^{\infty} N_j \epsilon_{t+j}, \tag{4}$$

where the coefficient matrices N_j decay to zero at a geometric rate as $j \to \infty$ and, when convenient, $N_j = 0$, $j < -(n-1)r$, will be assumed. This can be seen by writing $\Pi(z)^{-1} = (\det \Pi(z))^{-1} \Xi(z) = M(z)$, where $\Xi(z)$ is the adjoint polynomial matrix of $\Pi(z)$ with degree at most $(n-1)r$. Then, $\det \Pi(B)\, u_t = \Xi(B)\, \epsilon_t$ and, by the definition of u_t,

$$\Phi\left(B^{-1}\right) w_t = \Xi(B)\, \epsilon_t,$$

where $w_t = |\Pi(B)|\, y_t$. Now, one can find a $0 < \delta_2 < 1$ such that $\Phi\left(z^{-1}\right)^{-1} \Xi(z)$ has a well defined power series representation $\Phi\left(z^{-1}\right)^{-1} \Xi(z) = \sum_{j=-(n-1)r}^{\infty} N_j z^{-j} = N\left(z^{-1}\right)$ for $|z| > 1 - \delta_2$ (see Lanne and Saikkonen (2013)).

Hence, from (2) it follows that the process y_t itself has the representation

$$y_t = \sum_{j=-\infty}^{\infty} \Psi_j \epsilon_{t-j}, \tag{5}$$

where Ψ_j $(n \times n)$ is the coefficient matrix of z^j in the Laurent series expansion of $\Psi(z) \overset{def}{=} \Phi\left(z^{-1}\right)^{-1} \Pi(z)^{-1}$ which exists for $1 - \delta_2 < |z| < 1 + \delta_1$ with Ψ_j decaying to zero at a geometric rate as $|j| \to \infty$. The representation (5) implies that y_t is a stationary and ergodic process with finite second moments.

Taking conditonal expectation of equation (1) conditional on current and past values of y_t, it is seen that in the noncausal model, the elements of the Φ_j $(j = 1, \ldots, s)$ matrices capture the dependence of the variables included in y_t on their future expectated values. Alternatively, the conditional expectation of moving average representation (5),

$$y_t = \sum_{j=-\infty}^{s-1} \Psi_j E_t\left(\epsilon_{t-j}\right) + \sum_{j=s}^{\infty} \Psi_j \epsilon_{t-j}.$$

shows how noncausality implies dependence on future errors. This follows from the fact that, in the noncausal case y_t and ϵ_{t+j} are correlated, and consequently, $E_t\left(\epsilon_{t+j}\right) \neq 0$ for some $j \geq 0$. This also implies that future errors can be predicted by past values of the process y_t, which, in turn, can be interpreted as the errors containing factors not included in the model that are predictable by the variables in the VAR model (see Lanne and Saikkonen (2013) for a more elaborate discussion on this issue). Hence, the presence of noncausality might be seen symptomatic of missing variables whose effects are captured by the noncausal specification. This suggests that allowing for noncausality is likely to mitigate the effects of misspecification in VAR analysis. In addition to missing variables, misspecification of functional form may give rise to noncausality. As pointed out by Lanne and Saikkonen (2013), the noncausal VAR model has a nonlinear causal representation (see also Gourieroux and Zakoïan (2013) for a discussion on this point in the univariate noncausal AR model). While little is known of the implied form of nonlinearity, the noncausal VAR model can nevertheless be seen as a convenient shorthand representation of a complicated nonlinear process. Hence, a noncausal VAR model is likely to capture potential misspecification in the form of both missing variables or nonlinearity. The simulation results of Lof (2012) show that noncausality is easily confounded with very different econometric and economic nonlinear models (including the exponential smooth transition autoregression and financial models with heterogenous agents), lending support to these interpretations.

Because of the properties pointed out above, a potential application where the noncausal VAR model might prove useful, is testing for Granger causality that is known to depend on the variables included in the model (see, e.g., Lütkepohl 2005, 49–51)). Moreover,there is a growing interest in generalizing the Granger causality test to allow for nonlinearities (see, e.g., Péguin-Feissolle et al. (2013) and the references therein), and the noncausal vector autoregression seems to offer one relatively general nonlinear model that the test can be based on. Our approach to testing for Granger

causality is discussed more thoroughly in Section 5.3 in conjunction with the empirical application.

Finally, it should be pointed out that noncausal autoregressive models cannot be identified by second order properties or Gaussian likelihood. Therefore, meaningful application of the noncausal VAR model (1) requires that the error term ϵ_t is non-Gaussian. For details on the identifiablity of the noncausal VAR model and the assumptions needed for the derivation of the likelihood function we refer to Lanne and Saikkonen (2013). In this paper, we assume that the distribution of ϵ_t is multivariate t with scale matrix Σ and degrees of freedom λ.

3 Estimation

Lanne and Saikkonen (2013) studied maximum likelihood (ML) estimation of the noncausal VAR model (1). Our estimation method is built upon their work as well as our previous work on the Bayesian analysis of noncausal AR models (see Lanne, Luoma, and Luoto (2012)). In particular, our basic estimation algorithm is a straightforward extension of their Metropolis-within-Gibbs sampler (see also Geweke (2005, p. 206)). It is described in Subsection 3.2, and it exploits the fact that the full conditional posterior distributions of $\Pi_1, \ldots, \Pi_r$, $\Phi_1, \ldots, \Phi_s$, and Σ can be readily sampled from their known distributions. Our experience is that, in general, the sampler works well and convergence occurs rapidly.

In the general case ($r > 0$, $s > 0$), however, the posterior distribution of the parameters of (1) may be multimodal. If this occurs, the sampler tends to be inefficient and it may get stuck at one of the modes. It is important to note that this is not a serious problem if one is interested in forecasting, because the predictive distribution of the model in (1) turns out to be relatively invariant with respect to the multimodal posterior distribution (the forecasting procedure is described in Section 4). However, it is well known that multimodality complicates the estimation of the marginal likelihood and, if not properly handled, makes the commonly used approaches such as importance sampling and density ratio marginal likelihood approximation (see Gelfand and Dey (1994)) ill-suited for this purpose. Therefore, for the estimation of the marginal likelihood, we propose an alternative algorithm based on a Mixture of t by Importance Sampling weighted Expectation Maximization (MitISEM) algorithm of Hoogerheide, Opschoor, and van Dijk (2012). This algorithm is explained in Subsection 3.3.

3.1 Likelihood function

For the Bayesian analysis of the noncausal VAR model in (1), we need to derive the distribution of the observations conditional on the parameters, i.e., the likelihood function, and specify the prior distribution of the parameters. We start by describing the likelihood function, whose detailed derivation can be found in Lanne and Saikkonen (2013). The choice of the prior distribution is described in the next subsection. To simplify notation in our subsequent developments, we define the matrices Π and Φ, which are obtained by stacking Π_j' for $j = 1, ..., r$ and Φ_j' for $j = 1, ..., s$, respectively.

As mentioned in Section 2, we assume that ϵ_t follows the multivariate t distribution with scale matrix Σ and degrees of freedom λ. To make the model operational, we reparametrize ϵ_t in the following manner:

$$\epsilon_t = \widetilde{\omega}_t^{-\frac{1}{2}} \eta_t, \tag{6}$$

where η_t is a multivariate normally distributed random vector ($\eta_t \sim N(0, \Sigma)$), and $\lambda\widetilde{\omega}_t$ follows the chi-square distribution with λ degrees of freedom ($\lambda\widetilde{\omega}_t \sim \chi^2(\lambda)$). Under the chosen parameterization, y_t generated by (1) is conditionally Gaussian conditional on Σ and $\widetilde{\omega}_t$. As will be seen, this property is critical in building a decent posterior sampler (see also Geweke (1993), and Lanne, Luoma, and Luoto (2012)). Notice also that the random vector $(\widetilde{\omega}_1, \ldots, \widetilde{\omega}_T)$ can be interpreted as a vector of parameters with hierarchical priors $\lambda\widetilde{\omega}_t \sim \chi^2(\lambda)$ $(t = 1, \ldots, T)$ and $\lambda \sim Exp(\underline{\lambda})$, where $\underline{\lambda}$ is a prior hyperparameter.

The first step in the derivation of the likelihood function is writing the observed data $y = (y_1', ..., y_T')'$ in terms of vector $\boldsymbol{z} = (\boldsymbol{z}_1', \boldsymbol{z}_2', \boldsymbol{z}_3')'$, whose elements $\boldsymbol{z}_1 = (u_1', ..., u_r')'$, $\boldsymbol{z}_2 = \left(\epsilon_{r+1}', ..., \epsilon_{T-s}'\right)'$, and $\boldsymbol{z}_3 = (v_{1,T-s+1}', ..., v_{s,T}')'$, by (3) and (4), are independent. Here,

$$v_{k,T-s+k} = w_{T-s+k} - \sum_{j=-(n-1)r}^{-k} N_j \epsilon_{T-s+k+j}, \quad k = 1, ..., s, \tag{7}$$

and the sum is interpreted as zero when $k > (n-1)r$, that is, when the lower bound exceeds the upper bound. Note that, by (1) and (4), $v_{k,T-s+k}$ can be expressed as a function of the observed data $\boldsymbol{y}$ and that the representation $v_{k,T-s+k} = \sum_{j=-k+1}^{\infty} N_j \epsilon_{T-s+k+j}$ holds, showing that $v_{k,T-s+k}$ $(k = 1, ..., s)$ are indeed independent of ϵ_t, $t \leq T-s$. Thus, by (6) and the preceding discussion, the joint (conditional) density function of $\boldsymbol{z}$ conditional on $\widetilde{\boldsymbol{\omega}} = \left(\widetilde{\omega}_{r+1}, \ldots, \widetilde{\omega}_{T-s}\right)'$ and Σ can be expressed as

$$p(\boldsymbol{z} \,|\widetilde{\boldsymbol{\omega}}, \Sigma) = p(\boldsymbol{z}_1) \left(\prod_{t=r+1}^{T-s} p(\epsilon_t \,|\widetilde{\omega}_t, \Sigma) \right) p(\boldsymbol{z}_3), \tag{8}$$

where $p(\cdot)$ denotes a density function.

As shown in Section 3.1 of Lanne and Saikkonen (2013), the random vector $\boldsymbol{z}$ is related to the data vector $\boldsymbol{y} = (y_1', ..., y_T')'$ by a linear transformation of the form $\boldsymbol{z} = \boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{y}$, where $\boldsymbol{H}_1$, $\boldsymbol{H}_2$, and $\boldsymbol{H}_3$ are $T \times T$ nonsingular transformation matrices that depend on the parameters Π and Φ. Furthermore, the determinants of $\boldsymbol{H}_2$ and H_3 equal unity (for details of these matrices, see Lanne and Saikkonen (2013)). Thus, by (8), the conditional joint density function of the data $\boldsymbol{y}$ conditional on the parameters and $\widetilde{\boldsymbol{\omega}}$ can be expressed as

$$p(\boldsymbol{y}|\widetilde{\boldsymbol{\omega}}, \theta) = p(\boldsymbol{z}_1(\vartheta)) \left(\prod_{t=r+1}^{T-s} p\left(\Pi(B)\Phi\left(B^{-1}\right) y_t | \widetilde{\omega}_t, \Sigma\right) \right) p(\boldsymbol{z}_3(\vartheta)) |\boldsymbol{H}_1|. \tag{9}$$

In addition to the distinct elements of the matrix Σ, that is, the vector $\sigma = \text{vech}(\Sigma)$, the parameter vector θ also contains $\vartheta = (\pi', \phi')'$, where $\pi = \text{vec}(\Pi)$, and $\phi = \text{vec}(\Phi)$. The components of $\mathbf{z}$ can be expressed in terms of the observed data and parameters. Specifically, $\mathbf{z}_1(\vartheta)$ is defined by replacing u_t in the definition of $\mathbf{z}_1$ by $\Phi(B^{-1}) y_t$ $(t = 1, ..., r)$. Moreover, $\mathbf{z}_3(\vartheta)$ is defined similarly by replacing $v_{k,T-s+k}$ in the definition of $\mathbf{z}_3$ by an analog with $a(B) y_{T-s+k}$ and $\Pi(B)\Phi(B^{-1}) y_{T-s+k+j}$ used in place of w_{T-s+k} and $\epsilon_{T-s+k+j}$, respectively, where $j = -(n-1)r, ..., -k$, $k = 1, ..., s$, and

$$|\Pi(z)| = a(z) = 1 - a_1 z - \cdots - a_{nr} z^{nr}. \tag{10}$$

Lanne and Saikkonen (2013) also show that the determinant of H_1 is independent of the sample size T, and thus, following them, we propose to approximate the (conditional) joint density of $\boldsymbol{y}$ by the second factor of (9):

$$p(\boldsymbol{y}|\widetilde{\boldsymbol{\omega}}, \theta) \approx \prod_{t=r+1}^{T-s} p(\epsilon_t(\vartheta)|\widetilde{\omega}_t, \Sigma), \tag{11}$$

where

$$p(\epsilon_t(\vartheta)|\widetilde{\omega}_t, \Sigma) = \frac{\widetilde{\omega}_t^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \widetilde{\omega}_t \epsilon_t(\vartheta)' \Sigma^{-1} \epsilon_t(\vartheta) \right],$$

and

$$\epsilon_t(\vartheta) = u_t(\vartheta_2) - \Pi_1 u_{t-1}(\vartheta_2) - \cdots - \Pi_r u_{t-r}(\vartheta_2). \tag{12}$$

Finally, it should be pointed out that the approximate likelihood of Lanne and Saikkonen (2013) is obtained by multiplying (11) by the (hierarchical) prior density of $\widetilde{\boldsymbol{\omega}}$ and then integrating out $\widetilde{\boldsymbol{\omega}}$:

$$p(\boldsymbol{y}|\theta) = \int \prod_{t=r+1}^{T-s} p(\epsilon_t(\vartheta)|\widetilde{\omega}_t, \Sigma) p(\widetilde{\boldsymbol{\omega}}|\lambda) d\widetilde{\boldsymbol{\omega}}, \tag{13}$$

where, from the assumption $\lambda\widetilde{\omega}_t \sim \chi^2(\lambda)$, the quantity $p(\widetilde{\boldsymbol{\omega}}|\lambda)$ is obtained as the following product:

$$p(\widetilde{\omega}_t|\lambda) = \left[2^{\lambda/2}\Gamma(\lambda/2)\right]^{-1} \lambda^{\lambda/2}\widetilde{\omega}_t^{\lambda/2-1} \exp\left[-\lambda\widetilde{\omega}_t/2\right] \quad \text{for } t = r+1,\ldots,T-s. \quad (14)$$

A closed form representation of the approximate likelihood (13) is obtained from (8) by replacing the latent data expression $p(\epsilon_t|\widetilde{\omega}_t, \Sigma)$ with the density $p(\epsilon_t|\Sigma, \lambda)$, and using the change of variable theorem:

$$p(\boldsymbol{y}|\theta) = \prod_{t=r+1}^{T-s} p(\epsilon_t(\vartheta)|\Sigma, \lambda), \quad (15)$$

where we augment θ by the degrees of freedom parameter λ, and $p(\epsilon_t(\vartheta)|\Sigma, \lambda) = t_n(\epsilon_t(\vartheta)|\Sigma; \lambda)$, the density function of the multivariate t-distribution for an n-dimensional random vector $\epsilon_t(\vartheta)$ with zero mean, scale matrix Σ, and degrees of freedom λ.[1] For simplicity, we shall usually drop the word 'approximate' and refer to (11) as the likelihood function.

3.2 Basic Algorithm

We now turn to the estimation of the parameters of (1). As already discussed, this is accomplished by a multivariate generalization of the Metropolis-within-Gibbs sampler of Lanne, Luoma, and Luoto (2012). The detailed derivation of the full conditional posteriors exploited in the sampler are given in the appendix, while we here only describe the algorithm.

The conditional posteriors can be obtained from the product of (11), (14), and the joint prior density of Φ, Π, Σ, and λ. Following the literature, we assume the independent normal-Wishart prior for $\phi = \text{vec}(\Phi)$, $\pi = \text{vec}(\Pi)$, and Σ (see, e.g., Kadiyala and Karlsson (1997)), and, as already mentioned, an exponential prior for λ. In particular, $\pi \sim N(\underline{\pi}, \underline{V}_\pi) I(\pi)$, $\phi \sim N(\underline{\phi}, \underline{V}_\phi) I(\phi)$, $\Sigma \sim iW(\underline{S}, \underline{\nu})$, and $\lambda \sim Exp(\underline{\lambda})$, where iW is used to denote an inverse-Wishart distribution, and $\underline{\phi}$, $\underline{V}_\phi$, $\underline{\pi}$, $\underline{V}_\pi$, $\underline{S}$, $\underline{\nu}$, and $\underline{\lambda}$ are the prior hyperparameters assumed to be known by the researcher.

[1] The density function of the multivariate t-distribution for an n-dimensional random vector x with zero mean, λ degerees of freedom, and covariance matrix $\frac{\lambda}{\lambda-2}\Sigma$ is given by

$$t_n(x|\mu, \Sigma; \lambda) = \frac{\Gamma[(\lambda+n)/2]}{(\lambda\pi)^{n/2}\Gamma(\lambda/2)\sqrt{|\Sigma|}} \left(1 + \frac{1}{\lambda}x'\Sigma^{-1}x\right)^{-(\lambda+n)/2}$$

where $\Gamma(\cdot)$ is the gamma function and $\lambda > 2$ is assumed.

Indicator functions $I(\phi)$ and $I(\pi)$ equal unity in the stationary region defined by (2) and zero otherwise.

To simplify notation, we introduce a $Tn \times 1$ vector $\mathbf{y}^*$ and a $Tn \times sn^2$ matrix $\mathbf{X}^*$, which are obtained by stacking $y_t^* = \widetilde{\omega}_t^{1/2} \Pi(B) y_t$ and $X_t^* = \widetilde{\omega}_t^{1/2} \Pi(B) X_t$ for $t = r+1, \ldots, T-s$, where $X_t = I_n \otimes \left(y_{t+1}', \ldots, y_{t+s}'\right)$, respectively. We also define the matrices $\mathbf{Y}$ and $\mathbf{U}$, whose tth rows ($t = r+1, \ldots, T-s$) are given by $u_t^* = \widetilde{\omega}_t^{1/2} u_t'(\vartheta_2)$ and $U_t^* = \widetilde{\omega}_t^{1/2} \left(u_{t-1}'(\vartheta_2), \ldots, u_{t-r}'(\vartheta_2)\right)$, respectively. Then, the full conditional posterior distributions of ϕ, π, and Σ under the given prior distributions have the following expressions:

$$\begin{aligned}
\phi \left| \mathbf{y}, \pi, \Sigma, \widetilde{\boldsymbol{\omega}} \right. &\sim N\left(\overline{\phi}, \overline{V}_\phi\right) I(\phi), && (16)\\
\pi \left| \mathbf{y}, \phi, \Sigma, \widetilde{\boldsymbol{\omega}} \right. &\sim N\left(\overline{\pi}, \overline{V}_\pi\right) I(\pi), && (17)\\
\overline{V}_\phi^{-1} &= \underline{V}_\phi^{-1} + \mathbf{X}^{*\prime} \boldsymbol{\Omega} \mathbf{X}^*, \; \overline{\phi} = \overline{V}_\phi \left(\underline{V}_\phi^{-1} \underline{\phi} + \mathbf{X}^{*\prime} \boldsymbol{\Omega} \mathbf{y}^*\right), \\
\overline{V}_\pi^{-1} &= \underline{V}_\pi^{-1} + \Sigma^{-1} \otimes \mathbf{U}'\mathbf{U}, \; \overline{\pi} = \overline{V}_\pi \left(\underline{V}_\pi^{-1} \underline{\pi} + vec\left(\mathbf{U}'\mathbf{Y}\Sigma^{-1}\right)\right),
\end{aligned}$$

with $\boldsymbol{\Omega} = I_{T-r-s} \otimes \Sigma^{-1}$, and

$$\begin{aligned}
\Sigma \left| \mathbf{y}, \pi, \phi, \widetilde{\boldsymbol{\omega}} \right. &\sim iW\left(\overline{S}, \overline{\nu}\right), \; \overline{\nu} = \underline{\nu} + T - s - r, && (18)\\
\overline{S} &= \underline{S} + \mathbf{E}'\mathbf{E}, \; \mathbf{E} = \mathbf{Y} - \mathbf{U}\Pi.
\end{aligned}$$

The full conditional posterior distributions of the remaining parameters, $\widetilde{\boldsymbol{\omega}}$ and λ, can be sampled from

$$\left[\lambda + \epsilon_t(\vartheta)' \Sigma^{-1} \epsilon_t(\vartheta)\right] \widetilde{\omega}_t \left| \mathbf{y}, \pi, \phi, \Sigma, \lambda \right. \sim \chi^2(\lambda + n) \quad (t = r+1, \ldots, T-s), \qquad (19)$$

and, by a Metropolis within Gibbs step, from a distribution with the density kernel:

$$\begin{aligned}
p\left(\lambda \left| \mathbf{y}, \widetilde{\boldsymbol{\omega}} \right.\right) \propto & \left[2^{\lambda/2} \Gamma(\lambda/2)\right]^{-(T-r-s)} \lambda^{\lambda(T-r-s)/2} \left(\prod_{t=r+1}^{T-s} \widetilde{\omega}_t^{(\lambda-2)/2}\right) \\
& \times \exp\left[-\left(\frac{1}{\underline{\lambda}} + \frac{1}{2} \sum_{t=r+1}^{T-s} \widetilde{\omega}_t\right) \lambda\right]. \qquad (20)
\end{aligned}$$

Given the starting values of ϕ, π, Σ, $\widetilde{\boldsymbol{\omega}}$, and λ, the expressions in (16)–(20) are used sequentially to obtain an estimate of the posterior distribution of the paremeters. In particular, the first four expressions are standard and can be readily used to simulate random numbers. Note also that the stationarity restrictions (2) can be imposed by discarding the draws from the unresticted posterior distributions of ϕ and π that do not lie in the stationary region. Following Geweke (2005), we simulate from the conditional posterior of the degree-of-freedom parameter λ (20) using

an independence-chain MH algorithm. As a candidate distribution for λ we use the univariate normal distribution with mean equal to the mode of (20) and precision parameter equal to the negative of the second derivative of the log posterior density, evaluated at the mode. The acceptance probability is calculated using (20).

As already pointed out above, the sampler works well when the posterior distribution is unimodal. However, if the posterior is multimodal, it tends to be inefficient and may get stuck at one of the modes. For these cases, in Section 3.3 below, we propose an alternative algorithm based on a MitISEM algorithm of Hoogerheide, Opschoor, and van Dijk (2012) that we apply in the estimation of the marginal likelihood.

3.3 Marginal Likelihood Estimation

In the general case ($r > 0$, $s > 0$), because of the complexity of model (1), the marginal posterior distributions of its parameters tend to exhibit non-elliptical shapes such as skewness and multimodality. As is well known, the Gibbs sampler does not mix well with respect to a multimodal target posterior distribution, but tends to get stuck at one of the modes (subspaces). Therefore, in this subsection, we explain how to quickly construct an accurate approximation to the non-elliptical target posterior distribution. This approximation can then be used as a candidate density, say, in the Metropolis–Hastings algorithm or in importance sampling. In this paper, we use the latter to estimate the marginal likelihood of model (1) (see Geweke (2005, p. 257) for a detailed discussion).

As already mentioned, the proposed procedure closely resembles that of Hoogerheide, Opschoor, and van Dijk (2012), and we refer to their paper for a more detailed discussion on the topic (see also Cappé et al. (2008)). Following their recommendation, we use a mixture of multivariate t distributions as the candidate density:

$$f\left(\theta \left| \psi \right.\right) = \sum_{g=1}^{G} \alpha_g t_k \left(\theta \left| \mu_g, V_g; \nu_g \right.\right), \tag{21}$$

where $\psi = \left(\mu_1', \ldots, \mu_G', \text{ vech}\left(V_1\right)', \ldots, \text{vech}\left(V_G\right)', \nu_1, \ldots, \nu_G, \alpha_1, \ldots, \alpha_{G-1}\right)'$, the mixing probabilities α_g satisfy $\sum_{g=1}^{G} \alpha_g = 1$, and $t_k\left(\theta \left| \mu_g, V_g; \nu_g \right.\right)$ ($k = (s+r) \times n^2 + n \times (n+1)/2 + 1$) refers to the density function of the multivariate t distribution with mode μ_g, (positive definite) scale matrix V_g, and degrees of freedom ν_g. The number of mixture components G is determined iteratively as explained at the end of this subsection.

In order to obtain a convenient approximation to the target posterior density, we minimize the Kullback–Leibler divergence between the target and candidate distrib-

utions, $\int p(\theta|\mathbf{y}) \log \frac{p(\theta|y)}{f(\theta|\psi)} d\theta$, with respect to ψ. Because the elements of vector ψ do not enter the posterior density $p(\theta|y)$, this is equivalent to maximizing

$$\int [\log f(\theta|\psi)] p(\theta|\mathbf{y}) d\theta = E[\log f(\theta|\psi)], \tag{22}$$

where E is the expectation with respect to the posterior distribution $p(\theta|\mathbf{y})$.

We propose the following two-step procedure for computing the parameters ψ of the candidate mixture distribution (21). In the first stage, the basic algorithm described in the previous subsection is run several times, each time using very different starting values θ_0. All the simulated N_0 draws are then together used to approximate a sample from the posterior $p(\theta|\mathbf{y})$. An initial estimate ψ_0 can be found using the Expectation Maximization (EM) algorithm to maximize an estimate of $E[\log f(\theta|\psi)]$, given by

$$\frac{1}{N_0} \sum_{i=1}^{N_0} \log f\left(\theta^i|\psi\right). \tag{23}$$

In the second stage, we use the initial estimate ψ_0 to draw an independently and identically distributed sample θ^i $(i = 1, \ldots, N)$ from $f(\theta|\psi_0)$ *in* (21). From this sample we then calculate

$$\frac{1}{N} \sum_{i=1}^{N} W^i \log f\left(\theta^i|\psi\right) \quad \text{with } W^i = \frac{p\left(\theta^i|\mathbf{y}\right)}{f\left(\theta^i|\psi_0\right)}. \tag{24}$$

This is a simulation-consistent estimate of expression (22), which can seen by noting that

$$\begin{aligned} \int [\log f(\theta|\psi)] p(\theta|\mathbf{y}) d\theta &= \int \left[\frac{p(\theta|\mathbf{y})}{f(\theta|\psi_0)} \log f(\theta|\psi)\right] f(\theta|\psi_0) d\theta \\ &= E\left[\frac{p(\theta|\mathbf{y})}{f(\theta|\psi_0)} \log f(\theta|\psi)\right], \end{aligned}$$

Now, ψ can be found by maximizing (24) by the EM algorithm. Once the candidate density has been obtained, it is successfully used to estimate the marginal likelihood $p(\mathbf{y})$, and as mentioned above, to that end, we employ importance sampling.

Hoogerheide, Opschoor, and van Dijk (2012) use the EM algorithm to maximize (24) in their bottom-up procedure which iteratively adds components into the mixture (21), starting with one multivariate t distribution. Conversely, we start with a reasonably large number of distributions and remove the (nearly) singular ones (i.e., those with (nearly) singular covariance matrices and very small probability weights). This

can be done because our basic algorithm tends to converge rapidly to the subspace (mode) closest to the starting values, enabling us to quickly construct a reasonably good approximation to the posterior distribution (a few thousand draws of each θ_0 seems to be sufficient for the approximation). Hence, we only need to calculate the Importance Sampling (IS) weights W^i $(i = 1, \ldots, N)$ once, while in the MitISEM algorithm the IS weights must be evaluated at each iteration. Note that because the basic algorithm described in Section 3.2 tends to get stuck at the local mode, our procedure for initial estimation, is not able to move between different subspaces (modes) in a balanced fashion, that is, according to their posterior probabilities. This suggests that our initial estimates of the mixing probabilities α_g $(g = 1, \ldots, G)$ may be poor. However, in the empirical application in Section 5, we find it very hard to improve the accuracy of our final approximation by adding additional components in the mixture.

4 Forecasting

In this section, we consider evaluating the posterior predictive distribution of y_{T+h} $(h \geq 1)$ and, unless otherwise stated, we shall assume that the model is noncausal and multivariate, i.e., $s > 0$ and $n > 1$. Our starting point is equation (4), which is made operational by approximating the infinite sum on the right hand side by a finite sum. Recalling that w_t can be written as $w_t = |\Pi(B)| y_t = a(B) y_t$, where

$$|\Pi(z)| = a(z) = 1 - a_1 z - \cdots - a_{nr} z^{nr},$$

and substituting this into equation (4), we obtain the approximation

$$y_{T+h} \approx a_1 y_{T+h-1} + \cdots + a_{nr} y_{T+h-nr} + \sum_{j=-(n-1)r}^{M-h} N_j \epsilon_{T+h+j}. \tag{25}$$

M is a positive integer, and because the coefficient matrices N_j decay to zero at a geometric rate as $j \to \infty$, the approximation error can be made negligible by setting M sufficiently large. An approximate predictive distribution of y_{T+h} for $h > 0$, conditional on information in period T, can be computed recursively starting from $h = 1$, provided we are able to evaluate the conditional distribution of the last term on the right hand side of (25) for every $h > 0$. In the univariate case $(n = 1)$ considered by Lanne et al. (2012a,b) this term contains the errors $\epsilon_{T+1}, ..., \epsilon_{T+M}$ only, facilitating a straightforward way to obtain forecasts. However, as emphasized by Nyberg and Saikkonen (2013), in a multivariate case the error terms $\epsilon_{T+1-(n-1)r}, ..., \epsilon_T$ are also

involved, and because $\epsilon_{T-s+1}, ..., \epsilon_T$ $(s > 0)$ cannot be expressed as functions of the observed data (cf., (1)), additional complications arise.

The forecasting procedure is based on the joint density of the augmented data vector $\left(\mathbf{y}', \epsilon'_{T+1}, ..., \epsilon'_{T+M}\right)'$. The derivation of this density and the resulting sampling algorithm are described in the following two subsections, respectively. As the posterior predictive distribution tends to be relatively invariant with respect to a multimodal posterior distribution, the procedure is built upon the simpler algorithm described in Section 3.2.

4.1 Augmented Data Density

The conditional distribution of the last term on the right hand side of (25) is obtained by computing the joint density of the augmented data vector $\left(\mathbf{y}', \epsilon'_{T+1}, ..., \epsilon'_{T+M}\right)'$. A detailed derivation of this joint density can be found in Nyberg and Saikkonen (2013), and, in the univariate case, in Lanne, Luoma, and Luoto (2012), and Lanne, Luoto, and Saikkonen (2012). We first review the result of Nyberg and Saikkonen (2013) (modified slightly for ensuing derivations), and then use it to build an algorithm that yields a predictive distribution of y_{T+h} for $h > 0$ as a by-product.

Because the future errors enter the conditional density of y (9) only through the elements of $z_3 = (v'_{1,T-s+1}, ..., v'_{s,T})'$ (see the discussion following (9)), the conditional density of y and ϵ^+ can be written as

$$p\left(\boldsymbol{y}, \boldsymbol{\epsilon}^+ | \widetilde{\boldsymbol{\omega}}, \theta\right) = p(\boldsymbol{z}_1) \left|\boldsymbol{H}_1\right| \left(\prod_{t=r+1}^{T-s} p\left(\Pi\left(B\right) \Phi\left(B^{-1}\right) y_t \left| \widetilde{\omega}_t, \Sigma\right.\right) \right) p(\boldsymbol{z}_3\left(\vartheta\right), \boldsymbol{\epsilon}^+), \quad (26)$$

where $\epsilon^+ = \left(\epsilon'_{T+1}, ..., \epsilon'_{T+M}\right)$, and we remove the term $p(\boldsymbol{z}_1) \left|\boldsymbol{H}_1\right|$ from the above expression based on the discussion preceding (11). Thus, in order to find the (conditional) density of $(\mathbf{y}, \boldsymbol{\epsilon}^+)$, we need to find the joint density of $(\mathbf{z}_3, \epsilon^+)$, whose derivation, in turn, can be reduced to the derivation of the joint density function of the vector $(\boldsymbol{\zeta}'_1, \boldsymbol{\zeta}'_2)'$ defined below. Note that from (7), $\mathbf{z}_3 = (v'_{1,T-s+1}, ..., v'_{s,T})'$ with $v_{k,T-s+k} = \sum_{j=-k+1}^{\infty} N_j \epsilon_{T-s+k+j}$, $k = 1, ..., s$. Define the $sn \times n$ matrices

$$\boldsymbol{N}_j = \begin{bmatrix} N_j \\ \vdots \\ N_{j-s+1} \end{bmatrix}, \quad j = 0, 1, ...,$$

and write $\boldsymbol{z}_3$ as $\sum_{j=0}^{\infty} N_j \epsilon_{T-s+1+j}$. Define next the $sn \times 1$ vector $\zeta_1 = z_3 - \sum_{j=sn}^{\infty} N_j \epsilon_{T-s+j+1}$

and the $sn\,(n-1)\times 1$ vector $\zeta_2 = \left(\epsilon'_{T+1},\ldots,\epsilon'_{T-s+sn}\right)'$ and write

$$\left[\begin{array}{c}\boldsymbol{\zeta}_1\\ \boldsymbol{\zeta}_2\end{array}\right] = \left[\begin{array}{cc}\boldsymbol{Q}_{11} & \boldsymbol{Q}_{12}\\ \mathbf{0} & I_{sn(n-1)}\end{array}\right]\left[\begin{array}{c}\epsilon_{T-s+1}\\ \vdots\\ \epsilon_{T-s+sn}\end{array}\right] \equiv \boldsymbol{Q}\left[\begin{array}{c}\epsilon_{T-s+1}\\ \vdots\\ \epsilon_{T-s+sn}\end{array}\right], \tag{27}$$

where $\boldsymbol{Q}_{11} = [\boldsymbol{N}_0\ \cdots\ \boldsymbol{N}_{s-1}]$, $\boldsymbol{Q}_{12} = [\boldsymbol{N}_s\ \cdots\ \boldsymbol{N}_{sn-1}]$, and $\left(\epsilon'_{T-s+1},...,\epsilon'_{T-s+sn}\right)'$ is an $sn^2\times 1$ vector.

Now, it can readily be checked *that* (see Appendix A.2 in Nyberg and Saikkonen (2013))

$$p(\boldsymbol{z}_3,\boldsymbol{\epsilon}^+) \approx p(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2\,|\boldsymbol{\varepsilon}^+)\prod_{t=T-s+sn+1}^{T+M} p\left(\epsilon_t\right), \tag{28}$$

where $\boldsymbol{\varepsilon}^+ = \left(\epsilon'_{T-s+sn+1},\ldots,\epsilon'_{T+M}\right)$, and

$$\boldsymbol{\zeta}_1 = \boldsymbol{z}_3 - \sum_{j=sn}^{M+s-1}\boldsymbol{N}_j\epsilon_{T-s+j+1}.$$

Furthermore, from (27) it is seen that the matrix $\boldsymbol{R}=\boldsymbol{Q}^{-1}$ describes the linear transformation $(\boldsymbol{\zeta}_1',\boldsymbol{\zeta}_2')' \to (\epsilon_{T-s+1},\ldots,\epsilon_{T-s+sn})'$ so that

$$p(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2\,|\boldsymbol{\varepsilon}^+) = \prod_{t=T-s+1}^{T-s+sn} p\left(\epsilon_t\right)|\boldsymbol{R}|, \tag{29}$$

where

$$\left[\begin{array}{c}\epsilon_{T-s+1}\\ \vdots\\ \epsilon_{T-s+sn}\end{array}\right] = \left[\begin{array}{cc}\boldsymbol{Q}_{11}^{-1} & -\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\\ \mathbf{0} & I_{sn(n-1)}\end{array}\right]\left[\begin{array}{c}\boldsymbol{\zeta}_1\\ \boldsymbol{\zeta}_2\end{array}\right].$$

Thus, $\epsilon_{T-s+1},\ldots,\epsilon_T$ are obtained as functions of the observed data through $\boldsymbol{z}_3(\vartheta)$ and the errors $\boldsymbol{\epsilon}^+ = \left(\epsilon'_{T+1},...,\epsilon'_{T+M}\right)'$, and we (sometimes) use the notation $\epsilon_t(\vartheta,\boldsymbol{\epsilon}^+)$ for $t=T-s+1,\ldots,T$ to make the dependence explicit. By combining (26), (28), and (29), and using (6), we obtain the expression

$$p\left(\boldsymbol{y},\boldsymbol{\epsilon}^+\,|\widetilde{\boldsymbol{\omega}},\theta\right) \approx \prod_{t=r+1}^{T} p\left(\epsilon_t\left(\vartheta,\boldsymbol{\epsilon}^+\right)|\widetilde{\omega}_t,\Sigma\right)\prod_{t=T+1}^{T+M} p\left(\epsilon_t\,|\Sigma,\lambda\right)|\boldsymbol{R}|, \tag{30}$$

which is a multivariate extension of the expression (9) in Lanne, Luoma, and Luoto (2012).

4.2 Simulating the Predictive Distribution

Next, we turn to the estimation of the predictive distribution of y_{T+h} for $h > 0$. By exploiting the conditional density of y and ϵ^+ (30), this distribution can be obtained by a straightforward extension of the procedure of Lanne, Luoma, and Luoto (2012). This procedure involves estimating the posterior distributions of the parameters ϕ, π, Σ, and λ, and the latent variables $\widetilde{\boldsymbol{\omega}}$ and $\boldsymbol{\epsilon}^+$.

Using the priors described in Section 3.2, the full conditional posterior densities of ϕ, π, Σ, $\widetilde{\boldsymbol{\omega}}$, and λ can be obtained by following exactly the same steps as in estimating the posterior distribution of the parameters, as discussed in Section 3.2. However, because we now use the joint (conditional) density of y and ϵ^+ (30) instead of the (approximate) data density (11), we must multiply the conditional posterior density of each of the parameters ϕ, π, Σ, $\widetilde{\boldsymbol{\omega}}$, and λ given in Section 3.2 by the following expression:

$$|\boldsymbol{R}| \prod_{t=T-s+1}^{T} p\left(\epsilon_t\left(\vartheta, \boldsymbol{\epsilon}^+\right)|\widetilde{\omega}_t, \Sigma\right) \prod_{t=T+1}^{T+M} p\left(\epsilon_t|\Sigma, \lambda\right). \tag{31}$$

To see this, note that the augmented data density (30) can be expressed as the product of two terms, the (approximate) data density (11), and (31). It follows that we can use the components of the conditional posterior densities as proposal densities in the Metropolis-Hastings (MH) algorithm (see Chib and Greenberg (1994)), and by the definition of the Metropolis–Hastings algorithm, the acceptance probabilities of the algorithm given below are then based solely on (31).

Due to the high-dimensional posterior distribution of $\epsilon^+ = \left(\epsilon'_{T+1}, \ldots, \epsilon'_{T+M}\right)'$ (with nM parameters to be estimated in total), sampling all the error terms using a single-block proposal distribution leads to an inefficient sampler. Therefore, the error terms $\epsilon_{T+1}, \ldots, \epsilon_{T+M}$ are sampled using multiple blocks. In particular, sampling is performed by the randomized block MH method of Chib and Ramamurthy (2010), where at each iteration the error terms are first randomly clustered into an arbitrary number of blocks, and then simulated one block at time by a MH step. However, the performance of the sampler can be further increased by applying the randomized block MH method only for the first m error terms because typically only the first, say, $m < M$ error terms are strongly dependent on the data, and therefore their posterior distributions are potentially correlated. The remaining error terms can be conveniently grouped into additional b groups. In other words, the sampler can be tuned by the parameters m and b, and the hierarchical prior density $p\left(\epsilon_{T+1}|\Sigma, \lambda\right) \cdots p\left(\epsilon_{T+M}|\Sigma, \lambda\right)$ is used as a candidate density for each individual block.

We start from initial parameter values ϕ^0, π^0, ϵ^{+0}, Σ^0, $\widetilde{\boldsymbol{\omega}}^0$, and λ^0, and then sequentially simulate ϕ^j, π^j, $\boldsymbol{\epsilon}^{+j}$, Σ^j, $\widetilde{\boldsymbol{\omega}}^j$, and λ^j for $j = 1, \ldots, J$ by the following steps:

1. Draw a candidate ϕ^* from $\phi \left| \mathbf{y}, \pi^{j-1}, \Sigma^{j-1}, \widetilde{\boldsymbol{\omega}}^{j-1} \right. \sim N\left(\overline{\phi}, \overline{V}_\phi\right) I(\phi)$ (see (16)) and accept this proposal with probablity

$$\min\left\{\frac{\prod_{T-s+1}^{T} p\left(\epsilon_t\left(\phi^*, \pi^{j-1}, \boldsymbol{\epsilon}^{+,j-1}\right) \left| \widetilde{\omega}_t^{j-1}, \Sigma^{j-1}\right.\right) \left|\boldsymbol{R}\left(\phi^*, \pi^{j-1}\right)\right|}{\prod_{T-s+1}^{T} p\left(\epsilon_t\left(\vartheta^{j-1}, \boldsymbol{\epsilon}^{+,j-1}\right) \left| \widetilde{\omega}_t^{j-1}, \Sigma^{j-1}\right.\right) \left|\boldsymbol{R}\left(\vartheta^{j-1}\right)\right|}, 1\right\}$$

   If the proposed value is rejected, set ϕ^j at its current value ϕ^{j-1}. Here we use the notation $R(\phi, \pi)$ to indicate the dependence of R on ϕ and π (see (4) and (27), and the related discussions).

2. Draw a candidate π^* from $\pi \left| \mathbf{y}, \phi^j, \Sigma^{j-1}, \widetilde{\boldsymbol{\omega}}^{j-1} \right. \sim N\left(\overline{\pi}, \overline{V}_\pi\right) I(\pi)$ (see (17)) and accept the proposal with probability

$$\min\left\{\frac{\prod_{T-s+1}^{T} p\left(\epsilon_t\left(\phi^j, \pi^*, \boldsymbol{\epsilon}^{+,j-1}\right) \left| \widetilde{\omega}_t^{j-1}, \Sigma^{j-1}\right.\right) \left|\boldsymbol{R}\left(\phi^j, \pi^*\right)\right|}{\prod_{T-s+1}^{T} p\left(\epsilon_t\left(\phi^j, \pi^{j-1}, \boldsymbol{\epsilon}^{+,j-1}\right) \left| \widetilde{\omega}_t^{j-1}, \Sigma^{j-1}\right.\right) \left|\boldsymbol{R}\left(\phi^j, \pi^{j-1}\right)\right|}, 1\right\}$$

   If the proposed value is rejected, set π^j at its current value π^{j-1}.

3. Randomly group the errors $\epsilon_{T+1}^{j-1}, \ldots, \epsilon_{T+m}^{j-1}$ into the b_j blocks $\boldsymbol{\epsilon}_1^{+,j-1}, \boldsymbol{\epsilon}_2^{+,j-1}, ..., \boldsymbol{\epsilon}_{b_j}^{+,j-1}$.[2] *[HUOM!]* For $i = 1, \ldots, b_j + b$ draw a candidate block ϵ_i^{+*} using the hierarchical prior density $\prod_{t=T+1}^{T+M} p(\epsilon_t | \Sigma, \lambda)$, and accept the proposal with probability

$$\min\left\{\frac{\prod_{T-s+1}^{T} p\left(\epsilon_t\left(\vartheta^j, \boldsymbol{\epsilon}_i^{+*}, \boldsymbol{\epsilon}_{-i}^{+,j}\right) \left| \widetilde{\omega}_t^{j-1}, \Sigma^{j-1}\right.\right)}{\prod_{T-s+1}^{T} p\left(\epsilon_t\left(\vartheta^j, \boldsymbol{\epsilon}_i^{+,j-1}, \boldsymbol{\epsilon}_{-i}^{+,j}\right) \left| \widetilde{\omega}_t^{j-1}, \Sigma^{j-1}\right.\right)}, 1\right\}$$

   where $\boldsymbol{\epsilon}_{-i}^{+,j}$ contains the most currently updated values of all the error terms except for those in the ith block. If the proposal $\boldsymbol{\epsilon}_i^{+*}$ is rejected, set $\boldsymbol{\epsilon}_i^{+,j}$ at its current value $\boldsymbol{\epsilon}_i^{+,j-1}$.

[2] We follow the procedure of Chib and Ramamurthy (2010) to obtain the random blocks $\epsilon_1^+, \epsilon_2^+, \ldots, \epsilon_{b_j}^+$, in the jth iteration. The algorithm is started by randomly permuting $\epsilon_{T+1}, \ldots, \epsilon_{T+m}$. The shuffled error terms are denoted by $\epsilon_{T+\rho(1)}, \ldots, \epsilon_{T+\rho(m)}$, where $\rho(1), \ldots, \rho(m)$ is a permutation of the integers $1, 2, \ldots, m$. Then, the blocks $\epsilon_1^+, \epsilon_2^+, \ldots, \epsilon_{b_j}^+$ are obtained recursively as follows. The first block $\boldsymbol{\epsilon}_1^+$ is initialized at $\epsilon_{T+\rho(1)}$. Each error term $\epsilon_{T+\rho(l)}$ in turn for $l = 2, 3, \ldots, m$, is included in the first block with probability *(tuning parameter)* p_ϵ, and used to start a new block with probability $(1 - p_\epsilon)$. The procedure is repeated until each reshuffled error term is included in one of the blocks.

4. Draw a candidate Σ^* from $\Sigma\left|\mathbf{y},\pi^j,\phi^j,\boldsymbol{\epsilon}^{+j},\widetilde{\boldsymbol{\omega}}^{j-1}\right. \sim iW\left(\overline{S},\overline{\nu}\right)$ (see (18)) and accept the proposal with probability

$$\min\left\{\frac{\prod_{t=T+1}^{T+M} p\left(\epsilon_t\left|\Sigma^*,\lambda^{j-1}\right.\right)}{\prod_{t=T+1}^{T+M} p\left(\epsilon_t\left|\Sigma^{j-1},\lambda^{j-1}\right.\right)},1\right\}$$

If the proposed value is rejected, set Σ^j at its current value Σ^{j-1}. Here $\overline{S} = \underline{S}+\widetilde{\mathbf{E}}'\widetilde{\mathbf{E}}$, $\overline{\nu} = T-r+\underline{\nu}$, and $\widetilde{\mathbf{E}}$ is obtained by stacking $\epsilon_t'\left(\vartheta^j\right)$ for $t = r+1,\ldots,T-s$ and then $\epsilon_t'\left(\vartheta^j,\boldsymbol{\epsilon}^{+j}\right)$ for $t = T-s+1,\ldots,T$.

5. Draw $\widetilde{\omega}_t^j$ for $t = r+1,\ldots,T$ using (19):

$$\left[\lambda^{j-1}+\epsilon_t\left(\vartheta^j,\boldsymbol{\epsilon}^{+j}\right)'\left(\Sigma^j\right)^{-1}\epsilon_t\left(\vartheta^j,\boldsymbol{\epsilon}^{+j}\right)\right]\widetilde{\omega}_t\left|\mathbf{y},\pi^j,\phi^j,\boldsymbol{\epsilon}^{+j},\Sigma^j,\lambda^{j-1}\right. \sim \chi^2\left(\lambda^{j-1}+n\right).$$

6. Draw λ^j using an independence-chain MH algorithm. As a candidate distribution for λ, use a univariate normal distribution, with mean equal to the mode of (20) and precision parameter equal to the negative of the second derivative of the log posterior density evaluated at the mode. The acceptance probability can be calculated using the product of the right hand side of (20) and the product $\prod_{t=T+1}^{T+M} p\left(\epsilon_t\left|\Sigma,\lambda\right.\right)$.

At each iteration, calculate forecasts y_{T+h} for $h > 0$ $(h = 1,...H)$ using (25) with π^j,ϕ^j, and ϵ^{+j}. Note that the above procedure can be applied for the purely noncausal VAR ($r = 0$) model, by removing Step 2 from the algorithm and setting $\vartheta^j = \phi^j$.

5 Empirical Application

We illustrate the use of the noncausal Bayesian VAR model with an application to U.S. GDP growth and inflation for which a noncausal VAR model is found to provide superior fit and out-of-sample forecast performance over its causal counterpart. Moreover, we examine Granger causality (in distribution) between the two variables, and based on the selected noncausal specification, we find no evidence of Granger causality from inflation to GDP growth. As discussed in Section 5.3, this indicates that the output gap is not driving inflation in the new Keynesian model. However, there seems to be a reverse Granger causal relationship, not detected in the corresponding conventional causal VAR model.

Both series are computed as $400\ln\left(Z_t/Z_{t-1}\right)$, where Z_t is either the GDP or the implicit price deflator of the GDP. The resulting series are denoted by x_t and π_t,

respectively. Our quarterly data set runs from 1955:1 to 2013:2, and the source of the data is the FRED database of the Federal Reserve Bank of St. Louis.

In estimation, we use the priors discussed in Section 3.2 above. We set the prior hyper parameter $\underline{\lambda}$ at 5, because relatively small values of $\underline{\lambda}$ seem to increase the performance of the sampler. The VAR coeficients ϕ and π are assumed prior independent, and the elements of the hyperparameters $\underline{\phi}$ and $\underline{\pi}$ are set to zero. Following Litterman (1980, 1986), we set the diagonal elements of $\underline{V}_\pi$ and $\underline{V}_\phi$ such that the prior standard deviations of the parameters for own and foreign lags (or leads) equal γ_1/l^{γ_3} and $\sigma_i\gamma_1\gamma_2/\sigma_j l^{\gamma_3}$, respectively, where $l = 1,\ldots,r$ (or $l = 1,\ldots,s$). Here the ratio σ_i/σ_j accounts for the different units of measurement of the dependent variable $(i = 1,\ldots,n)$ and jth $(j \neq i)$ explanatory variable. The parameter $\gamma_1 > 0$ is often referred to as the overall tightness of the prior, $0 < \gamma_2 \leq 1$ as the relative tightness of the other variables, and $\gamma_3 > 0$ as the lag decay rate. The values of these hyperparameters are set at $\gamma_1 = 2$, $\gamma_2 = 1$, and $\gamma_3 = 1$, and, following the literature, σ_i^2 is set at the residual standard error of a univariate causal AR(p) $(p = r+s)$ model for variable i $(i = 1,\ldots,n)$. The degrees of freedom parameter $\underline{\nu}$ is set to 10. We also assume that the prior scale matrix of Σ is diagonal. In particular, $\underline{S} = (\underline{\nu} - n - 1)\text{diag}(\sigma_1^2,\ldots,\sigma_n^2)$, indicating that $E(\Sigma) = \text{diag}(\sigma_1^2,\ldots,\sigma_n^2)$. With these parameter values, independent normal-Wishart prior is relatively flat.

5.1 Estimation Results

We estimate all causal and noncausal second, third and fourth-order VAR models and compute their marginal likelihoods. As discussed in Section 3.3 above, the marginal likelihoods are estimated using importance sampling (see, e.g., Geweke (2005, p. 257)). In the general case ($r > 0$, $s > 0$), the importance density function (21) is obtained by the procedure explained in Section 3.3.[3] Throughout, the results are based on $N = 100,000$ independent draws from (21). The resulting mixture importance distributions typically involve three component distributions, two of which have modes that are relatively far apart (the detailed results, not reported, are available upon request.).[4]

[3]Note that, in the purely causal and noncausal cases, we use $t_k\left(\theta \,\middle|\, \widehat{E(\theta\,|\mathbf{y})}, \widehat{var(\theta\,|\mathbf{y})}; 20\right)$ as the importance density function. Here $\widehat{E(\theta\,|\mathbf{y})}$ and $\widehat{var(\theta\,|\mathbf{y})}$ refer to the estimates of $E(\theta\,|\mathbf{y})$ and $\text{var}(\theta\,|\mathbf{y})$, respectively, calculated from the posterior distribution of θ, obtained by the algorithm of Section 3.2.

[4]To obtain an initial estimate ψ_0 (cf. (23)) for the parameters of the mixture importance density (21), the basic algorithm explained in Section 3.2 is run 15 times, each time using very different

The estimated log marginal likelihoods of all estimated models and their numerical standard errors (obtained by the Delta method) are presented in Table 1. There is clear evidence in favor of noncausality, and hence nonlinearity, as conditional on the order, a noncausal model with one lag always maximizes the marginal likelihood, while the causal model yields the smallest value. Among all models, the noncausal VAR(1,2) model is the best, followed by the second-order VAR(1,1) model. Noncausality has previously been found in U.S. inflation series by Lanne and Saikkonen (2011), and Lanne, Luoma, and Luoto (2012), among others. Very small standard errors indicate accurate estimation, and hence, facilitate model selection. The error distribution indeed seems to be fat-tailed, as required for identification; the posterior mode of the degree-of-freedom parameter λ equals 4.19. For comparison, we also computed the maximum likelihood estimates of Lanne and Saikkonen (2013), and the posterior modes of all parameters turned out to lie close to the maximum likelihood estimates. However, the posterior densities of most coefficients are bimodal, with one clearly dominating mode.[5]

5.2 Forecasts

As discussed in Section 4, predictive distributions are obtained as a by-product of the estimation of the noncausal VAR model. In order to gauge the forecast performance, we compute pseudo out-of-sample forecasts from a number of models for the period 1970:1 to 2013:2. The forecasts are computed recursively, at each step reestimating each model using an expanding data window starting at 1955:1. We consider the forecast horizons of one, four, and eight quarters, as is common in the inflation and GDP growth forecasting literature.

We report the results of two evaluation criteria, the root mean squared forecast error (RMSFE) based on the median of the predictive distribution and the sum of log predictive likelihoods (PL) computed over the forecast period are reported. The RMSFE summarizes the accuracy of point forecasts, while the PL yields information on the forecasting performance of the entire predictive density. Geweke (1999, 2001), Geweke and Amisano (2010), and Chan et al. (2013), among others have also used the latter metric to evaluate the accuracy of density forecasts, and as emphasized by Geweke (1999, 2001), it is also very closely connected to the log of the marginal likelihood. As a matter of fact, when the log predictive likelihoods are evaluated at

randomly selected starting values (N_0 is set at 15×5000; see the discussion following (23)).

[5]Detailed estimation results are not presented to save space, but they are available upon request.

the observed values over the entire sample period (1955:1–2013:2 here), these two measures are equal. Following Bauwens et al. (2011) and Clark and Doh (2011), we compute the predictive likelihoods using kernel density estimation of the forecasted densities of the VAR(r, s) models.

The sums of log predictive likelihoods of all third-order VAR models are reported in Table 2, and the VAR(1,2) model selected in the in-sample analysis above, outperforms the other specifications by a wide margin.at all forecast horizons. The corresponding figures for the univariate density forecasts reported in the right-hand side panel of Table 3 also indicate the superiority of the VAR(1,2) model in predicting inflation. For GDP growth, the results are not quite as clear cut: the VAR(1,2) model is the winner at the one-quarter horizon, but it is beaten by the VAR(0,3) model at the two longer forecast horizons.

As far as the point forecasts are concerned, the result in the left-hand side panel of Table 3 show that for inflation the purely noncausal VAR(0,3) model is the most accurate at the four and eight-quarter horizons, while it is marginally outperformed by the VAR(2,1) model at the one one-quarter horizon. Also for GDP growth the noncausal models always outperform the causal VAR(3,0) model. However, at the one and eight-quarter horizons, it is the VAR(2,1) model that yields the most accurate point forecasts, with the VAR(1,2) and VAR(0,3) models being the winners at the four-quarter horizon.

Probably the most surprising finding is that the univariate AR(1,2) model yields more accurate point and density forecasts of GDP growth than any of the VAR models, indicating that inflation contains no useful information for future GDP growth over and above the univariate noncausal model. Moreover, the AR(1,2) model outperforms the causal AR(3) model (not shown), attesting to the ability of the noncausal model to take effects of missing variables (other than inflation) into account. In contrast, for inflation the univariate AR model is clearly inferior to any of the VAR models, which suggests that GDP growth is useful in forecasting inflation in ways not captured by the univariate noncausal model.

5.3 Granger Causality

Because the noncausal VAR model can capture effects of missing variables and misspecified functional form, it is particularly useful in checking for Granger causality, as discussed above. Moreover, this can be done in a straightforward manner by means of Bayesian analysis, while conducting the corresponding classical test seems compli-

cated, or potentially even impossible. This follows from the difficulty of expressing the hypothesis of no marginal predictive power in terms of the parameters of the VAR model (see Nyberg and Saikkonen (2013, Section 4.1) for further discussion). Our Bayesian approach simply relies on comparing the marginal predictive likelihoods of the univariate and bivariate models to check whether the variable excluded from the univariate model has marginal predictive power for the other variable (at any forecast horizon). Hence, our procedure is based on comparing out-of-sample predictions (at all horizons), in line with Granger's (1969) seminal paper. In practice, this comparison is conducted at a number of forecast horizons to confirm the robustness of the findings. While the Granger noncausality test is typically defined in terms of the mean squared forecast error, our procedure corresponds to the concept of Granger causality in distribution defined by Droumaguet and Woźniak (2012), who propose a similar approach. Instead of predictive likelihoods, however, these authors use in-sample marginal likelihoods, which is feasible in their Markov-switching VAR model, allowing for the imposition of the Granger causality restrictions.

As pointed out at the beginning of this section, testing for Granger causality from inflation to GDP growth is particularly interesting because it can also be seen as a test of the new Keynesian model. In particular, assuming GPD growth is a reasonable proxy for the marginal cost, inflation should Granger cause it if the marginal cost indeed is driving inflation in accordance with the model (see, e.g., Rudd and Whelan (2005) and the references therein). In Table 4, we report the results of Granger causality analysis only for the forecast horizons of one, four and eight quarters, but the conclusions are robust with respect to horizon. Twice the logarithmic Bayes factors of the bivariate VAR(1,2) model against the univariate AR(1,2) model for GDP growth are also negative at all prediction horizons considered, indicating virtually no predictive ability of inflation for GDP growth over and above ist own history.[6] Supposing GDP growth is a reasonable proxy for the marginal cost, this can be interpreted as evidence against the new Keynesian model as it indicates that marginal cost is not driving inflation. Interestingly, however, there is strong evidence in favor of Granger causality from GDP growth to inflation, with twice the logarithmic Bayes factors around 20. In view of the results in Subsection 5.2, these findings are not surprising. The former outcome is also common in the previous literature (see, e.g., Lanne and Luoto (2013) and the references therein), whereas there is little evidence

[6] According to Kass and Raftery (1995), values less than 2 of twice the natural logarithm of the Bayes factor indicate no evidence, while values greater than 20 indicate very strong evidence in favor of Granger causality.

in favor of Granger causality from GDP growth to inflation, and based on the causal VAR(3,0) model, also we were not able to find Granger causality in either direction.[7]

6 Conclusion

In this paper, we have devised Bayesian methods of estimation and forecasting in the noncausal VAR model. In particular, we have proposed a relatively fast and reliable posterior simulator that yields the predictive distribution as a by-product. It is well known, however, that the posterior distributions of the parameters of nonlinear models tend to exhibit non-elliptical shapes such as skewness and multimodality, and based on our empirical findings, the noncausal VAR model is not an exception. Fortunately, it turned out that this has only marginal effect on the predictive distribution, but it nevertheless complicates the estimation of the marginal likelihood. Therefore, to successfully estimate the marginal likelihood of the model, we also proposed an alternative estimation procedure that closely resembles the MitISEM algorithm of Hoogerheide, Opschoor, and van Dijk (2012).

We demonstrated the new methods with an empirical application to U.S. inflation and GDP growth for which a noncausal VAR model turned out to be superior in both in-sample and out-of-sample performance over its conventional causal counterpart. In addition, we found GDP growth to have predictive power for the future distribution of inflation, but not vice versa, which may be interpreted as evidence against the new Keynesian model, provided GDP growth is a reasonable proxy of the marginal cost. In contrast, in line with the previous literature, we found no Granger causality in either direction in the causal VAR model. This suggests that either Granger causality is nonlinear, and hence, not detected in the linear causal VAR model, or alternatively, the noncausal model is capable of capturing the effects of variables not included in the model in a way that facilitates detecting the Granger causal relationship from GDP growth to inflation. Of course, both of these factors may also be present simultaneously.

We have only applied our method to a low-dimensional vector autoregression. However, the method can be readily used for larger dimensions, such as a VAR model comprising of the seven variables in the US macroeconomic model of Smets and

[7] In the causal VAR(3,0) model, Granger causality can easily be checked by comparing the unrestricted model to the restricted model with the lags of the other variable set to zero in each equation in turn (cf. Droumaguet and Woźniak (2012)). In both cases, twice the logarithmic Bayes factor based on the in-sample sum of log marginal likelihoods (from 1955:1–2013:2) is less than unity.

Wouters (2007), but this kind of an exercise calls for an informative prior distribution that shrinks the parameters towards the chosen prior mean, hence preventing overfitting. Nevertheless, with larger models, overparameterization may cause convergence problems for our Metropolis-within-Gibbs sampler that tends to increase posterior correlation between the coefficients of leads and lags.

Appendix

In this appendix we derive the full conditional posterior distributions of the groups of unobservables, namely, ϕ, π, Σ, $\widetilde{\boldsymbol{\omega}}$, and λ, given in expressions (16), (17), (18), (19), and (20).

To derive the full conditional posterior of ϕ in the general case ($r > 0$, $s > 0$), note that the term $\Pi(B)\Phi(B^{-1})y_t$ in (9), and, hence, in (11) can be rewritten as $\Pi(B)[y_t - \Phi_1 y_{t+1} - \cdots - \Phi_s y_{t+s}] = \Pi(B)[y_t - X_t\phi]$, where $X_t = I_n \otimes \left(y'_{t+1}, \ldots, y'_{t+s}\right)$. From (11), this yields

$$\begin{aligned} p(\mathbf{y}\,|\theta, \widetilde{\boldsymbol{\omega}}) &\propto \exp\left[-\frac{1}{2}\sum_{t=r+1}^{T-s}(y_t^* - X_t^*\phi)'\,\Sigma^{-1}(y_t^* - X_t^*\phi)\right] \\ &= \exp\left[-\frac{1}{2}(\mathbf{y}^* - \mathbf{X}^*\phi)'\,\boldsymbol{\Omega}(\mathbf{y}^* - \mathbf{X}^*\phi)\right], \end{aligned} \tag{.32}$$

where $\mathbf{y}^*$ and $\mathbf{X}^*$ are obtained by stacking $y_t^* = \widetilde{\omega}_t^{1/2}\Pi(B)y_t$ and $X_t^* = \widetilde{\omega}_t^{1/2}\Pi(B)X_t$ for $t = r+1, \ldots, T-s$, respectively, and $\Omega = I_{T-r-s} \otimes \Sigma^{-1}$. Rewriting the right hand side of (.32) in terms of the generalized least squares estimator of ϕ (see, e.g., Hamilton (1994, p. 220)), multiplying the resulting equation by the appropriate prior density, and completing the square for ϕ, we obtain the prior distribution (16) in Section 3.2:

$$\phi\,|\mathbf{y}, \pi, \Sigma, \widetilde{\boldsymbol{\omega}} \sim N\left(\overline{\phi}, \overline{V}_\phi\right) I(\phi),$$

where $\overline{V}_\phi = \left(\underline{V}_\phi^{-1} + \mathbf{X}^{*\prime}\boldsymbol{\Omega}\mathbf{X}^*\right)^{-1}$ and $\overline{\phi} = \overline{V}_\phi\left(\underline{V}_\phi^{-1}\underline{\phi} + \mathbf{X}^{*\prime}\boldsymbol{\Omega}\mathbf{y}^*\right)$.Following Chib and Greenberg (1994), we draw from this full conditonal posterior by sampling from the untruncated variant, $N\left(\overline{\phi}, \overline{V}_\phi\right)$, until we obtain a draw that lies in the stationary region.

It is also straightforward to confirm that the full conditional posterior distribution of π under the chosen prior is

$$\pi\,|\mathbf{y}, \phi, \Sigma, \widetilde{\boldsymbol{\omega}} \sim N\left(\overline{\pi}, \overline{V}_\pi\right) I(\pi),$$

where $\overline{V}_\pi = \left(\underline{V}_\pi^{-1} + \Sigma^{-1} \otimes \mathbf{U}'\mathbf{U}\right)^{-1}$ and $\overline{\pi} = \overline{V}_\pi\left(\underline{V}_\pi^{-1}\underline{\pi} + vec\left(\mathbf{U}'\mathbf{Y}\Sigma^{-1}\right)\right)$. The tth rows of the matrices Y and U [vai Y and U?] are given by $u_t^* = \widetilde{\omega}_t^{1/2}u_t'(\vartheta_2)$ and

$U_t^* = \widetilde{\omega}_t^{1/2}\left(u'_{t-1}(\vartheta_2), \ldots, u'_{t-r}(\vartheta_2)\right)$, respectively. To see this, note that by (12), the conditional density in (11) can also be rewritten in terms of $u_t(\vartheta_2)$. Following Zellner (1971, p. 225), from (11), we then obtain

$$p(\mathbf{y}|\theta, \widetilde{\boldsymbol{\omega}}) \propto \exp\left[-\frac{1}{2} tr\left(\mathbf{Y} - \mathbf{U}\Pi\right)'\left(\mathbf{Y} - \mathbf{U}\Pi\right)\Sigma^{-1}\right], \tag{.33}$$

whose right hand side can be written in terms of ordinary least squares quantities. Thus, by the prior density of π and (.33), expression (17) in Section 3.2 *is* obtained by standard calculations (see, e.g., Zellner (1971, p. 240) and Karlsson (2012)). Again, exactly as in the case of ϕ, sampling from (17) is carried out by drawing from the untruncated variant, $N\left(\overline{\pi}, \overline{V}_\pi\right)$ until obtaining a draw that lies in the stationary region.

Note that, in the purely causal case ($s = 0$), the vectors u_t^* and U_t^* reduce to $\widetilde{\omega}_t^{1/2} y'_t$ and $\widetilde{\omega}_t^{1/2}\left(y'_{t-1}, \ldots, y'_{t-r}\right)'$, respectively, and the conditional posterior distribution of π is $\pi|\mathbf{y}, \Sigma, \widetilde{\boldsymbol{\omega}} \sim N\left(\overline{\pi}, \overline{V}_\pi\right) I(\pi)$, where $\overline{\pi}$ and $\overline{V}_\pi$ are as in (17). Furthermore, in the purely non-causal case ($r = 0$), the vectors u_t^* and U_t^* in (17), can be replaced with $\widetilde{\omega}_t^{1/2} y'_t$ and $\widetilde{\omega}_t^{1/2}\left(y'_{t+1}, \ldots, y'_{t+s}\right)$, respectively, and then the right hand side of (.33) can be expressed as $\exp\left[-\frac{1}{2} tr\left(\mathbf{Y} - \mathbf{U}\Phi\right)'\left(\mathbf{Y} - \mathbf{U}\Phi\right)\Sigma^{-1}\right]$. In this case, the full conditional posterior of ϕ reduces to $\phi|\mathbf{y}, \Sigma, \widetilde{\boldsymbol{\omega}} \sim N\left(\overline{\phi}, \overline{V}_\phi\right) I(\phi)$, where $\overline{V}_\phi = \left(\underline{V}_\phi^{-1} + \Sigma^{-1} \otimes \mathbf{U}'\mathbf{U}\right)^{-1}$ and $\overline{\phi} = \overline{V}_\phi\left(\underline{V}_\phi^{-1}\underline{\phi} + vec\left(\mathbf{U}'\mathbf{Y}\Sigma^{-1}\right)\right)$.

From (.33), it can also be seen that under the inverse-Wishart prior, Σ has the conditional posterior of the form

$$\Sigma|\mathbf{y}, \pi, \phi, \widetilde{\boldsymbol{\omega}} \sim iW\left(\overline{S}, \overline{\nu}\right),$$

where $\overline{S} = \underline{S} + \mathbf{E}'\mathbf{E}$, $\mathbf{E} = \mathbf{Y} - U\Pi$, and $\overline{\nu} = \underline{\nu} + T - s - r$. That is, the expression (18) in Section 3.2 follows directly from (.33).

The kernel of $p\left(\widetilde{\boldsymbol{\omega}}|\mathbf{y}, \pi, \phi, \Sigma, \lambda\right)$ is obtained as a product of (11) and (14), where $\widetilde{\omega}_{r+1}, \ldots, \widetilde{\omega}_{T-s}$ are conditionally independent. Specifically,

$$\begin{aligned} p\left(\widetilde{\omega}_t|\mathbf{y}, \pi, \phi, \Sigma, \lambda\right) \quad &\propto \quad \widetilde{\omega}_t^{\frac{n+\lambda}{2}-1} \\ &\qquad \exp\left[-\widetilde{\omega}_t\left(\lambda + \epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta)\right)/2\right]. \quad (t = r+1, \ldots, T-s) \end{aligned}$$

Thus, by the properties of the chi-squared distribution, we obtain expression (19):

$$\left[\lambda + \epsilon_t(\vartheta)'\Sigma^{-1}\epsilon_t(\vartheta)\right]\widetilde{\omega}_t|\mathbf{y}, \pi, \phi, \Sigma, \lambda \sim \chi^2(\lambda + n). \qquad (t = r+1, \ldots, T-s)$$

From (14) and the assumption $\lambda \sim Exp\left(\underline{\lambda}\right)$, it follows that the conditional posterior density of λ can be written as

$$p\left(\lambda \left| \mathbf{y}, \widetilde{\boldsymbol{\omega}} \right.\right) \propto \left[2^{\lambda/2}\Gamma\left(\lambda/2\right)\right]^{-(T-r-s)} \lambda^{\lambda(T-r-s)/2} \left(\prod_{t=r+1}^{T-s} \widetilde{\omega}_t^{(\lambda-2)/2}\right) \exp\left[-\left(\frac{1}{\underline{\lambda}} + \frac{1}{2}\sum_{t=r+1}^{T-s} \widetilde{\omega}_t\right)\lambda\right].$$

It is the hierarchical prior structure in which λ affects the data only through $\widetilde{\boldsymbol{\omega}}$ that lies behind this result.

Following Geweke (2005), we simulate from the conditional posterior of the degree-of-freedom parameter λ using an independence-chain MH algorithm. As a candidate distribution of λ, we use the univariate normal distribution with mean equal to the mode of (20) and precision parameter equal to the negative of the second derivative of the log posterior density, evaluated at the mode. The acceptance probability is determined by (20).

References

Bauwens L., G. Koop, D. Korobilis, and J.V.K. Rombouts (2011). A comparison of forecasting procedures for macroeconomic series: The contribution of structural break models. SIRE Discussion Papers 2011-25, Scottish Institute for Research in Economics (SIRE).

Cappé, O., R. Douc, A. Guillin, J.M. Marin, and C.P. Robert (2008). Adaptive importance sampling in general mixture classes. Statistics and Computing 18, 447–459.

Chan, J., G. Koop, and S. M. Potter (2013). A new model of trend in‡ation. Journal of Business & Economic Statistics 31, 94–106.

Chib, S., and S. Ramamurthy (2010). Tailored randomized block MCMC methods with application to DSGE models. Journal of Econometrics 155, 19–38.

Chib, S., and E. Greenberg (1994). Bayes inference in regression models with ARMA (p,q) errors. Journal of Econometrics 64, 183–206.

Clark, T.E., and T. Doh (2011). A Bayesian evaluation of alternative models of trend in‡ation. Working Paper 1134. Federal Reserve Bank of Cleveland.

Davis, R.A., and L. Song (2010). Noncausal Vector AR processes with application to financial time series. Working Paper. Columbia University.

DelNegro, M., and F. Schorfheide (2011). Bayesian macroeconometrics, in J. Geweke, G. Goop, and H van Dijk (eds.), Oxford Handbook of Bayesian Econometrics, Oxford University Press, 293–389.

Droumaguet, M., and T. Woźniak (2012). Bayesian testing of Granger causality in Markov-switching VARs. EUI Working Papers ECO 2012/06.

Gelfand A., and D. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. Journal of The Royal Statistical Society Series B 56, 501–514.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. Journal of Applied Econometrics 8, 19–40.

Geweke, J. (1999). Using simulation methods for Bayesian econometrics models: Inference, development and communication. Econometric Reviews, 18 1–73.

Geweke, J. (2001). Bayesian econometrics and forecasting. Journal of Econometrics 100, 11–15.

Geweke, J. (2005). Contemporary Bayesian Econometrics and Statistics. Wiley, Hoboken, New Jersey.

Geweke, J., and G. Amisano (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. International Journal of Forecasting 26, 216–230.

Gourieroux, C., and J.-M. Zakoïan (2013). Explosive bubble modelling by noncausal processes. Working Paper 2013-04. Centre de Recherche en Economie et Statistique.

Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438.

Hoogerheide, L., A. Opschoor, and H.K. van Dijk (2012). A class of adaptive importance sampling weighted EM algorithms for efficient and robust posterior and predictive simulation. Journal of Econometrics 171, 101-120.

Kadiyala, K.R., and S. Karlsson (1997). Numerical Methods for Estimation and Inference in Bayesian VAR-Models. Journal of Applied Econometrics 12, 99-132.

Karlsson, S. (2013). Forecasting with Bayesian vector autoregressions, in G. Elliott and A. Timmermann (eds), Handbook of Economic Forecasting, Vol 2, Elsevier, Amsterdam.

Kass, R.E., and A.E. Raftery (1995). Bayes factors. Journal of the American Statistical Association 90, 773–795.

Kohn, R. Asymptotic estimation and hypothesis testing results for vector linear time series models. Econometrica 47, 1005–1029.

Lanne, M., A. Luoma, and J. Luoto (2012). Bayesian model selection and forecasting in noncausal autoregressive models. Journal of Applied Econometrics 7, 812–830.

Lanne, M., and J. Luoto (2013). Does output gap, labour's share or unemployment rate drive inflation? Oxford Bulletin of Economics and Statistics (forthcoming).

Lanne, M., J. Luoto, and P. Saikkonen (2012). Optimal forecasting of noncausal autoregressive time series. International Journal of Forecasting 28, 623–631.

Lanne, M., and P. Saikkonen (2011). Noncausal autoregressions for economic time series. Journal of Time Series Econometrics 3 (3), Article 3.

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

Lof, M. (2012). Noncausality and asset pricing. Studies in Nonlinear Dynamics and Econometrics 17, 211–220.

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

Nyberg, H., and P. Saikkonen (2013). Forecasting with a noncausal VAR model. Computational Statistics and Data Analysis (forthcoming).

Péguin-Feissolle, A., B. Strikholm, and T. Teräsvirta (2013). Testing the Granger noncausality hypothesis in stationary nonlinear models of unknown

functional form. Communications in Statistics - Simulation and Computation 42, 1063–1087.

Rudd, J., and K. Whelan (2005). Does labor's share drive inflation? Journal of Money, Credit, and Banking 37, 297–312.

Table 1: Model selection.

Model	ln ML	Std.err.
VAR(2,0)	–941.10	0.0035
VAR(1,1)	–939.96	0.0066
VAR(0,2)	–940.07	0.0055
VAR(3,0)	–944.77	0.0045
VAR(2,1)	–942.45	0.0071
VAR(1,2)	–939.44	0.0045
VAR(0,3)	–941.21	0.0056
VAR(4,0)	–950.11	0.0089
VAR(3,1)	–947.47	0.0280
VAR(2,2)	–944.55	0.0192
VAR(1,3)	–943.77	0.0090
VAR(0,4)	–945.79	0.0056

The figures in the second and third columns are the sums of the logarithmic marginal likelihoods of all second, third and fourth-order VAR models for inflation and output growth from 1955:1 to 2013:2, and their standard errors, respectively.

Table 2: Sums of h-step-ahead log predictive likelihoods.

Model	$h = 1$	$h = 4$	$h = 8$
VAR(3,0)	–702.8	–770.8	–799.0
VAR(2,1)	–701.5	–773.1	–803.4
VAR(1,2)	–698.2	–763.9	–794.6
VAR(0,3)	–708.7	–767.5	–796.9

The figures are the sums of the log predictive likelihoods (ln PL(h)) with one, four and eight quarter forecast horizons (h) for each model. The forecasts are computed recursively in the period 1970:1–2013:2, at each step reestimating each model using an expanding data window starting at 1955:1.

Table 3: Pseudo out-of-sample forecast analysis.

	RMSFE			ln PL(h)		
Model	$h=1$	$h=4$	$h=8$	$h=1$	$h=4$	$h=8$
			Inflation			
VAR(3,0)	1.108	1.541	2.038	–258.2	–314.2	–355.3
VAR(2,1)	1.103	1.572	2.114	–259.1	–314.6	–356.1
VAR(1,2)	1.126	1.525	2.035	–253.8	–306.4	–347.3
VAR(0,3)	1.105	1.511	2.007	–261.0	–312.6	–352.4
AR(1,2)	1.131	1.654	2.166	–276.7	–328.9	–363.8
			GDP Growth			
VAR(3,0)	3.428	3.691	3.608	–449.2	–457.0	–446.7
VAR(2,1)	3.281	3.661	3.514	–447.7	–458.1	–447.1
VAR(1,2)	3.331	3.623	3.578	–445.7	–458.0	–449.4
VAR(0,3)	3.316	3.623	3.534	–449.7	–456.0	–445.5
AR(1,2)	3.188	3.404	3.362	–442.6	–448.6	–437.3

The figures are the root mean square forecast errors (RMSFE) and sums of the log predictive likelihoods (ln PL(h)) with one, four and eight quarter forecast horizons (h) for inflation and GDP growth. The forecasts are computed recursively in the period 1970:1–2013:2, at each step reestimating each model using an expanding data window starting at 1955:1.

Table 4: Granger causality analysis.

AR(1,2) Model for	$h = 1$	$h = 4$	$h = 8$
Inflation	19.9	19.5	14.33
GDP Growth	–2.7	–8.2	–10.5

The figures are twice the natural logarithm of the Bayes factor of the bivariate VAR(1,2) model against the univariate AR(1,2) for either inflation or GDP growth. The Bayes factors are are based on the sum of the h-step (marginal) log predictive likelihoods summed over 1970:1–2013:2.

Research Papers 2013

2013-42: Torben G. Andersen and Oleg Bondarenko: Reflecting on the VPN Dispute

2013-43: Torben G. Andersen and Oleg Bondarenko: Assessing Measures of Order Flow Toxicity via Perfect Trade Classification

2013-44: Federico Carlini and Paolo Santucci de Magistris: On the identification of fractionally cointegrated VAR models with the F(d) condition

2013-45: Peter Christoffersen, Du Du and Redouane Elkamhi: Rare Disasters and Credit Market Puzzles

2013-46: Peter Christoffersen, Kris Jacobs, Xisong Jin and Hugues Langlois: Dynamic Diversification in Corporate Credit

2013-47: Peter Christoffersen, Mathieu Fournier and Kris Jacobs: The Factor Structure in Equity Options

2013-48: Peter Christoffersen, Ruslan Goyenko, Kris Jacobs and Mehdi Karoui: Illiquidity Premia in the Equity Options Market

2013-49: Peter Christoffersen, Vihang R. Errunza, Kris Jacobs and Xisong Jin: Correlation Dynamics and International Diversification Benefits

2013-50: Georgios Effraimidis and Christian M. Dahl: Nonparametric Estimation of Cumulative Incidence Functions for Competing Risks Data with Missing Cause of Failure

2013-51: Mehmet Caner and Anders Bredahl Kock: Oracle Inequalities for Convex Loss Functions with Non-Linear Targets

2013-52: Torben G. Andersen, Oleg Bondarenko, Viktor Todorov and George Tauchen: The Fine Structure of Equity-Index Option Dynamics

2014-01 Manuel Lukas and Eric Hillebrand: Bagging Weak Predictors

2014-02: Barbara Annicchiarico, Anna Rita Bennato and Emilio Zanetti Chini: 150 Years of Italian CO2 Emissions and Economic Growth

2014-03: Paul Catani, Timo Teräsvirta and Meiqun Yin: A Lagrange Multiplier Test for Testing the Adequacy of the Constant Conditional Correlation GARCH Model

2014-04: Timo Teräsvirta and Yukai Yang: Linearity and Misspecification Tests for Vector Smooth Transition Regression Models

2014-05: Kris Boudt, Sébastien Laurent, Asger Lunde and Rogier Quaedvlieg: Positive Semidefinite Integrated Covariance Estimation, Factorizations and Asynchronicity

2014-06: Debopam Bhattacharya, Shin Kanaya and Margaret Stevens: Are University Admissions Academically Fair?

2014-07: Markku Lanne and Jani Luoto: Noncausal Bayesian Vector Autoregression