NONCAUSAL BAYESIAN VECTOR AUTOREGRESSION

MARKKU LANNE[a,b] AND JANI LUOTO[a*]

[a] *Department of Political and Economic Studies, University of Helsinki, Finland*

[b] *CREATES, Aarhus University, Denmark*

SUMMARY

We consider Bayesian analysis of the noncausal vector autoregressive model that is capable of capturing nonlinearities and effects of missing variables. Specifically, we devise a fast and reliable posterior simulator that yields the predictive distribution as a by-product. We apply the methods to postwar U.S. inflation and GDP growth. The noncausal model is found superior in terms of both in-sample fit and out-of-sample forecasting performance over its conventional causal counterpart. Economic shocks based on the noncausal model turn out highly anticipated in advance. We also find the GDP growth to have predictive power for future inflation, but not vice versa.

Keywords: Noncausal time series, non-Gaussian time series, Bayesian analysis, inflation, New Keynesian model.

* Correspondence to: Jani Luoto, Department of Political and Economic Studies, University of Helsinki, P.O.Box 17 (Arkadiankatu 7), FIN–00014 University of Helsinki, Finland. Phone +358-(0)294-1911, fax +358-(0)294-28736. 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 (2012), Lanne and Saikkonen (2013), and Gourieroux and Jasiak (2014). While these specifications differ somewhat from each other, they are all 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 complicates interpretation as the errors of the noncausal VAR model are 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 noncausal VAR model has certain benefits. First, its having a nonlinear causal representation indicates that is able to capture nonlinearities although the form of nonlinearity afforded by it is, in general, unknown. Moreover, it is capable of incorporating effects of missing variables, and it may, therefore, be useful in many macroeconomic and financial applications, where assessing the adequacy of the included set of variables can be problematic. In particular, the model can accommodate the effects of variables that are included in the agents' information set but not observed by the econometrician, and whose omission may give rise to nonfundamentalness (see Alessi et al. (2008) for a recent survey). A case in point is news concerning future economic developments, such as a tax increase that affects agents' behavior, but is not observable to the econometrician.

In this paper, we consider Bayesian analysis, including estimation, model selection, and forecasting of the noncausal VAR model of Lanne and Saikkonen (2013). 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). We also extend our new sampler such that it conveniently yields the posterior predictive distribution as a by-product. It is worth noting that the posterior distribution of the parameters of the model may be multimodal. This may be a problem for computing the marginal likelihood, and, therefore, we also devise an alternative sampler based on the Importance Sampling weighted Expectations Mazimization algorithm of Hoogerheide et al. (2012).

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 to its causal counterpart in point and density forecasting. Moreover, our Bayesian procedure finds no evidence of Granger causality from inflation to GDP growth. Provided GDP growth is a reasonable proxy for the marginal cost, this suggests that it is not driving inflation as the new Keynesian theory would imply. This finding is potentially even stronger than that obtained in the previous literature as it indicates the absence of even nonlinear Granger causality.

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 $(t = 0, \pm 1, \pm 2, ...)$ proposed by Lanne and Saikkonen (2013) is generated by

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

where $\Pi(B) = I_n - \Pi_1 B - \cdots - \Pi_r B^r$ $(n \times n)$ and $\Phi\left(B^{-1}\right) = I_n - \Phi_1 B^{-1} - \cdots - \Phi_s B^{-s}$ $(n \times n)$ are matrix polynomials in the backward shift operator B, and ϵ_t $(n \times 1)$ is a sequence of independent, identically distributed (continuous) random vectors with zero mean and finite positive definite covariance matrix. If $\Phi_j \neq 0$ 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 $\Pi(z)$ and $\Phi(z)$ $(z \in \mathbb{C})$ have their zeros outside the unit disc, i.e.,

$$\det \Pi(z) \neq 0, \quad |z| \leq 1, \quad \text{and} \quad \det \Phi(z) \neq 0, \quad |z| \leq 1. \tag{2}$$

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 $\Phi(z^{-1})^{-1}\, \Xi(z)$ has a well defined power series representation $\Phi(z^{-1})^{-1}\, \Xi(z) = \sum_{j=-(n-1)r}^{\infty} N_j z^{-j} = N(z^{-1})$ 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 $\Psi(z) \overset{def}{=} \Phi(z^{-1})^{-1}\, \Pi(z)^{-1}$ 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, potentially mitigating 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 the implied form of nonlinearity is, in general, unknown, the noncausal VAR model can nevertheless be seen as a convenient shorthand representation of a complicated nonlinear process. 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 this interpretation.

Finally, it should be pointed out that noncausal autoregressive models cannot be identified by second order properties or Gaussian likelihood. In other words, under Gaussianity, the maximum of the likelihood functions of the causal and noncausal VAR(p) models is the same. Supposing the noncausal VAR model is correct, its causal counterpart is misspecified with uncorrelated but not independent errors that can be predicted by past values of the series. This predictability shows up as the difference between the likelihood functions of the causal and noncausal models. Under Gaussianity, independence coincides with uncorrelatedness, so that there is no difference in the values of the maximized likelihood functions. 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. The t-distribution provides a convenient alternative for capturing fat tails prevalent in economic applications, and it has recently been found adequate in much of the empirical research employing univiariate noncausal autoregressive models cited above.

3. ESTIMATION AND INFERENCE

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 multivariate 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. This 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 $\mathbf{\Pi}$ and $\mathbf{\Phi}$, 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 $\boldsymbol{y} = (y_1', ..., y_T')'$ 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 being 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 $\widetilde{\boldsymbol{\omega}} = (\widetilde{\omega}_{r+1}, \ldots, \widetilde{\omega}_{T-s})'$ 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 $nT \times nT$ nonsingular transformation matrices that depend on the parameters $\mathbf{\Pi}$ and $\mathbf{\Phi}$. Furthermore, the determinants of $\boldsymbol{H}_2$ and $\boldsymbol{H}_3$ 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

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

In addition to the distinct elements of the matrix Σ, that is, the vector $\sigma=\text{vech}(\Sigma)$, the parameter vector θ also contains $\vartheta=(\vartheta_1,\vartheta_2)=(\pi',\phi')'$, where $\pi=\text{vec}(\mathbf{\Pi})$, and $\phi=\text{vec}(\mathbf{\Phi})$. The components of $\boldsymbol{z}$ can be expressed in terms of the observed data and the parameters. Specifically, $\boldsymbol{z}_1\left(\vartheta\right)$ is defined by replacing u_t in the definition of $\boldsymbol{z}_1$ by $\Phi\left(B^{-1}\right)y_t$ $(t=1,...,r)$. Moreover, $\boldsymbol{z}_3\left(\vartheta\right)$ is defined similarly by replacing $v_{k,T-s+k}$ in the definition of $\boldsymbol{z}_3$ by an analog with $a\left(B\right)y_{T-s+k}$ and $\Pi\left(B\right)\Phi\left(B^{-1}\right)y_{T-s+k+j}$ used in place of w_{T-s+k} and $\epsilon_{T-s+k+j}$, respectively, where $j=-\left(n-1\right)r,....,-k$, $k=1,...,s$, and

$$\left|\Pi\left(z\right)\right|=a\left(z\right)=1-a_1z-\cdots-a_{nr}z^{nr}. \quad (10)$$

It is important to realize that the quantities $p(z_1\left(\vartheta\right))$ and $p(z_3\left(\vartheta\right))$ specify the densities of the initial values $y_1,\ldots,y_r$ and the post sample observations $y_{T-s+1},\ldots,y_T$, respectively. Lanne and Saikkonen (2013) also show that the determinant of $\boldsymbol{H}_1$ is independent of the sample size T, and thus, following them, we propose to approximate the (conditional) joint density of $\boldsymbol{y}$ by the third factor of (9):

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

where

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

and

$$\epsilon_t\left(\vartheta\right)=u_t\left(\vartheta_2\right)-\Pi_1u_{t-1}\left(\vartheta_2\right)-\cdots-\Pi_ru_{t-r}\left(\vartheta_2\right). \quad (12)$$

Expression (11) is the exact conditional likelihood of $\boldsymbol{y}$ conditional on $y_1, \ldots, y_r$ and $y_{T-s+1}, \ldots, y_T$ (and $\widetilde{\boldsymbol{\omega}}$).[1]

3.2. Basic Algorithm

We now turn to the estimation of the parameters of model (1). As already discussed, this is accomplished by a multivariate generalization of the Metropolis-within-Gibbs sampler of Lanne, Luoma, and Luoto (2012). The full conditional posteriors exploited in the sampler can be obtained by routine calculations, given a few data transformations that allow for convenient conditioning. In the following, we first provide the required data tranfromations, and then briefly describe the algorithm.[2]

The conditional posteriors can be derived from the product of (11), the density of $\widetilde{\boldsymbol{\omega}}$ (see the discussion following (6)), and the joint prior density of $\boldsymbol{\Phi}$, $\boldsymbol{\Pi}$, Σ, and λ. Following the literature, we assume the independent normal-Wishart prior for $\pi = \text{vec}(\boldsymbol{\Pi})$, and Σ (see, e.g., Kadiyala and Karlsson (1997)), and, as already mentioned, an exponential prior for λ. In the same vein, a normal prior distribution is assumed for $\phi = \text{vec}(\boldsymbol{\Phi})$. In particular, $\pi \sim N\left(\underline{\pi}, \underline{V}_{\pi}\right) I\left(\pi\right)$, $\phi \sim N\left(\underline{\phi}, \underline{V}_{\phi}\right) I\left(\phi\right)$, $\Sigma \sim iW\left(\underline{S}, \underline{\nu}\right)$, and $\lambda \sim Exp\left(\underline{\lambda}\right)$, 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. Indicator functions $I\left(\phi\right)$ and $I\left(\pi\right)$ equal unity in the stationary region defined by (2) and zero otherwise.

To simplify notation, we introduce a $Tn \times 1$ vector $\boldsymbol{y}^*$ and a $Tn\ \times sn^2$ matrix $\boldsymbol{X}^*$, which are obtained by stacking $y_t^* = \widetilde{\omega}_t^{1/2} \Pi\left(B\right) y_t$ and $X_t^* = \widetilde{\omega}_t^{1/2} \Pi\left(B\right) X_t$ 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

[1] We thank a referee for pointing this out.

[2] A detailed derivation of the full conditional posteriors exploited in the sampler is available in the online Appendix.

$\boldsymbol{Y}$ and $\boldsymbol{U}$, 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:

$$\begin{aligned} \phi \,|\boldsymbol{y}, \pi, \Sigma, \widetilde{\boldsymbol{\omega}} \quad &\sim \quad N\left(\overline{\phi}, \overline{V}_\phi\right) I(\phi), & (13) \\ \pi \,|\boldsymbol{y}, \phi, \Sigma, \widetilde{\boldsymbol{\omega}} \quad &\sim \quad N\left(\overline{\pi}, \overline{V}_\pi\right) I(\pi), & (14) \\ \overline{V}_\phi^{-1} \quad &= \quad \underline{V}_\phi^{-1} + \boldsymbol{X}^{*\prime} \boldsymbol{\Omega} \boldsymbol{X}^*,\ \overline{\phi} = \overline{V}_\phi \left(\underline{V}_\phi^{-1} \underline{\phi} + \boldsymbol{X}^{*\prime} \boldsymbol{\Omega} \boldsymbol{Y}^* \right), \\ \overline{V}_\pi^{-1} \quad &= \quad \underline{V}_\pi^{-1} + \Sigma^{-1} \otimes \boldsymbol{U}' \boldsymbol{U},\ \overline{\pi} = \overline{V}_\pi \left(\underline{V}_\pi^{-1} \underline{\pi} + vec\left(\boldsymbol{U}' \boldsymbol{Y} \Sigma^{-1} \right) \right), \end{aligned}$$

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

$$\begin{aligned} \Sigma \,|\boldsymbol{y}, \pi, \phi, \widetilde{\boldsymbol{\omega}} \quad &\sim \quad iW\left(\overline{S}, \overline{\nu}\right),\ \overline{\nu} = \underline{\nu} + T - s - r, & (15) \\ \overline{S} \quad &= \quad \underline{S} + \boldsymbol{E}' \boldsymbol{E},\ \boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{U} \Pi. \end{aligned}$$

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

$$\left[\lambda + \epsilon_t (\vartheta)' \Sigma^{-1} \epsilon_t (\vartheta) \right] \widetilde{\omega}_t \,|\mathbf{y}, \pi, \phi, \Sigma, \lambda \sim \chi^2 (\lambda + n) \quad (t = r+1, \ldots, T-s), \qquad (16)$$

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

$$\begin{aligned} p(\lambda \,|\boldsymbol{y}, \widetilde{\boldsymbol{\omega}}) \quad &\propto \quad \left[2^{\lambda/2} \Gamma(\lambda/2) \right]^{-(T-r-s)} \lambda^{\lambda(T-r-s)/2} \left(\prod_{t=r+1}^{T-s} \widetilde{\omega}_t^{(\lambda-2)/2} \right) \\ &\quad \times \exp\left[- \left(\frac{1}{\underline{\lambda}} + \frac{1}{2} \sum_{t=r+1}^{T-s} \widetilde{\omega}_t \right) \lambda \right]. & (17) \end{aligned}$$

Given the starting values of ϕ, π, Σ, $\widetilde{\boldsymbol{\omega}}$, and λ, the expressions in (13)–(17) 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. Following Geweke (2005), we simulate from the conditional posterior of the degree-of-freedom parameter λ (17) using an independence-chain Metropolis-Hastings (MH) algorithm. As a candidate distribution for λ we use the univariate normal distribution with mean equal to the mode of (17) 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 (17).

As 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 efficienctly 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 G multivariate t distributions as the candidate density:

$$f\left(\theta\left|\psi\right.\right) = \sum_{g=1}^{G} \alpha_g t_k\left(\theta\left|\mu_g, V_g; \nu_g\right.\right), \tag{18}$$

where $\psi = \left(\mu_1', \ldots, \mu_G', \ \text{vech}\,(V_1)', \ldots, \text{vech}\,(V_G)', \nu_1, \ldots, \nu_G, \alpha_1, \ldots, \alpha_{G-1}\right)'$, 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.[3] 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 distributions, $\int p\left(\theta \left|\boldsymbol{y}\right.\right) \log \frac{p(\theta|y)}{f(\theta|\psi)} d\theta$, with respect to ψ. Because the elements of vector ψ do not enter the posterior density $p\left(\theta \left|\boldsymbol{y}\right.\right)$, this is equivalent to maximizing

$$\int \left[\log f\left(\theta \left|\psi\right.\right)\right] p\left(\theta \left|\boldsymbol{y}\right.\right) d\theta = E\left[\log f\left(\theta \left|\psi\right.\right)\right], \tag{19}$$

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

We propose the following two-step procedure for computing the parameters ψ of the candidate mixture distribution (18). 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, resulting in a large matrix comprising N_0 simulated draws that are then together used to approximate a sample from the posterior $p\left(\theta \left|\boldsymbol{y}\right.\right)$. An initial estimate ψ_0 can be found using the Expectation Maximization (EM) algorithm to maximize an estimate of $E\left[\log f\left(\theta \left|\psi\right.\right)\right]$, given by

$$\frac{1}{N_0} \sum_{i=1}^{N_0} \log f\left(\theta^i \left|\psi\right.\right). \tag{20}$$

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

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 (18). From this sample we then calculate

$$\frac{1}{N}\sum_{i=1}^{N} W^i \log f\left(\theta^i\,|\psi\right) \quad \text{with } W^i = \frac{p\left(\theta^i\,|\mathbf{y}\right)}{f\left(\theta^i\,|\psi_0\right)}. \tag{21}$$

This is a simulation-consistent estimate of expression (19), which can seen by noting that

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

Now, ψ can be found by maximizing (21) by the EM algorithm. Once the candidate density has been obtained, it is successfully used to estimate the marginal likelihood $p(\boldsymbol{y})$, and as mentioned above, to that end, we employ importance sampling.

Hoogerheide, Opschoor, and van Dijk (2012) use the EM algorithm to maximize (21) in their bottom-up procedure which iteratively adds components into the mixture (18), 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 the basic algorithm of Section 3.2 used to obtain initial estimates, tends to get stuck at the local mode, and hence is not able to move between different subspaces (modes) in a balanced fashion, i.e., 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, according to our experience, this hardly affects the quality of the final estimates, and in the empirical application in Section 5, we indeed do find it very hard to improve the accuracy of our final approximation by adding additional components in the mixture.

4. FORECASTING

As pointed out in the Introduction, the approach of Lanne, Luoma and Luoto (2012) can be extended to 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

$$y_{T+h} \approx a_1 y_{T+h-1} + \cdots + a_{nr} y_{T+h-nr} + \sum_{j=-(n-1)r}^{M-h} N_j \epsilon_{T+h+j}. \tag{22}$$

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. The 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 (22) 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 the 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 distribution of the augmented data vector $\left(\boldsymbol{y}', \epsilon'_{T+1}, ..., \epsilon'_{T+M}\right)'$. The derivation of this density and the resulting sampling algorithm are described in the online Appendix. Due to the high-dimensional joint posterior distribution of θ and $\epsilon_{T+1}, ..., \epsilon_{T+M}$, the procedure introduced in Subsection 3.3 is not computationally feasible for forecasting, and, therefore, the proposed method is built upon the simpler algorithm described in Section 3.2. As a matter of fact, the algorithm only needs to be expanded by one additional Gibbs step for $\epsilon_{T+1}, ..., \epsilon_{T+M}$.

It is important to note that if the posterior distribution of θ is in fact multimodal and the proposed sampler is not able to move between the different subspaces in a balanced fashion, some aspects of the 'true' predictive distribution may be lost. However, it is shown in Lanne and Saikkonen (2013) that the limiting distribution of the maximum likelihood estimator of the parameters of the noncausal VAR model is multivariate normal, indicating that multimodality is related to small sample sizes. Furthermore, it is our experience that commonly used informative Minnesota priors result in posterior distributions that are more easily handled by our algorithm.

5. NONCAUSAL VAR FOR U.S. GDP GROWTH AND INFLATION

We apply the noncausal Bayesian VAR model to the key U.S. macroeconomic variables, namely GDP growth and inflation. Both series are computed as $400 \ln\left(Z_t/Z_{t-1}\right)$, where Z_t denotes either the GDP or the implicit price deflator of the GDP. The resulting series are denoted by x_t and π_t, respectively. Both series are seasonally adjusted. 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. The VAR coefficients ϕ and π are assumed prior independent, and the elements of the hyperparameters $\underline{\phi}$ and $\underline{\pi}$ are set to zero. Following Litterman (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 the jth ($j \neq i$) explanatory variable, 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 $i = 1, \ldots, n$. 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 $\gamma_3 = 1$. We set $\underline{S} = (\underline{\nu} - n - 1)\text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, and the degree-of-freedom parameter $\underline{\nu}$ is set to 10. Finally, we set the prior hyper parameter $\underline{\lambda}$ at 5.

5.1. Estimation Results

We estimate all causal and noncausal second, third and fourth-order VAR models and compute their marginal likelihoods. They are estimated using importance sampling, and in the general case ($r > 0$, $s > 0$), the importance density function (18) is obtained by the procedure explained in Section 3.3. Throughout, the results are based on $N = 100,000$ independent draws from (18). 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.).

The log marginal likelihoods of all estimated models and their numerical standard errors (obtained by the delta method) are presented in Table I. 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 selected. The 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 marginal posterior distributions of most coefficients are multimodal, with one clearly dominating mode.

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 forecast performance, we compute pseudo out-of-sample forecasts from a number of models for the period 1970:1 to 2013:2. They 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. The RMSFE and PL summarize the accuracy of point and density forecasts, respectively. 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 II. The VAR(1,2) model selected in the in-sample analysis, outperforms the other specifications by a wide margin at the one and eight-quarter forecast horizons, while the VAR(0,3) model is the most accurate at the four-quarter horizon. The corresponding figures for the univariate density forecasts reported in the right-hand side panel of Table III also indicate the superiority of the VAR(0,3) and VAR(1,2) models.

As far as the point forecasts are concerned, the result in the left-hand side panel of Table III 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(0,3) models being the winner 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.

We finally check the forecasting results using some informative priors.[4] The conclusion that the noncausal VAR models are superior to their conventional causal counterparts in terms of point and density forecasting performance remains intact irrespective of the priors used. However, while the informative priors have negligible effect on the forecasting performance of the VAR models for inflation, they bring about substantial improvements in density and point forecast accuracy of GDP growth.

5.3. Granger Causality and Impulse Response Analysis

Because the noncausal VAR model can capture effects of missing variables and omitted nonlinearities, it is likely to alleviate the well-known dependence of Granger causality on the employed specification (see, e.g., Lütkepohl 2005, 49–51)). Bayesian analysis also lends itself to checking Granger causality in a straightforward manner, while conducting the corresponding classical test seems complicated, 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 noncausal VAR model (see Nyberg and Saikkonen (2013, Section 4.1) for further discussion).

[4] We consider the commonly used values of $\gamma_1 = \gamma_2 = 0.2$ (and $\gamma_1 = 0.6$ and $\gamma_2 = 0.1$). In addition, we check the results by setting the elements of $\underline{\phi}$ (or $\underline{\pi}$, in the purely causal case (s=0)) corresponding to the first leads (lags) of inflation and GDP growth to 0.8 and to 0.3, respectively, and the other elements of $\underline{\phi}$ and $\underline{\pi}$ to zero. Thus, for $s > 0$, we are assuming that the persistence in the U.S. inflation results from forward-looking behavior rather than dependence on past inflation. The detailed results are not presented to save space, but they are available upon request.

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). In practice, this comparison is conducted at a sufficiently large 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.

Assuming the 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 new Keynesian theory (see, e.g., Rudd and Whelan (2005) and the references therein). We computed the differences in marginal predictive likelihoods between the VAR(1,2) and the univariate AR(1,2) models at forecast horizons of one, four and eight quarters for each variable in turn. In the case of the GDP growth, these figures are negative at all prediction horizons considered, indicating virtually no predictive ability of inflation for the GDP growth over and above its own history. Interestingly, there is strong evidence in favor of the reverse Granger causality from the GDP growth to inflation, with the difference in the PLs around 10. ~~Whereas, there~~ is little evidence in the previous literature in favor of Granger causality from the GDP growth to inflation in data including the period after the mid-1980s (see, e.g., Lanne and Luoto (2013) and the references therein), and based on the causal VAR(3,0) model, also we were unable to find Granger causality in either direction.[5]

[5] 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, the Bayes Factor (BF), calculated as the ratio of the marginal likelihood of the restricted model to that of the unrestriced model, is

The noncausal VAR model also lends itself to impulse response analysis, with the elements of the Ψ_j matrix of its moving-average representation (5) being the impulse responses of the components of error ϵ_t at lag (or lead) j. For $j \leq -1$, the elements of Ψ_j, reveal to what extent each error is anticipated at each lead. Such anticipation effects reflect the information incorporated in the errors that helps the agents to predict future shocks.

To obtain one-standard-deviation orthogonal shocks, following Song and Davis (2012), we use a lower-triangular Choleski decomposition of the covariance matrix of the errors. We order inflation first, so the shock to the GDP growth has no *unexpected* immediate impact on inflation. This is in contrast to the corresponding recursive causal model, where *any* immediate impact of the shock to the GDP growth on inflation is precluded. Both shocks may still have an *expected* immediate impact on both variables. The estimated impulse response functions (not shown to save space) indicate that the variables are significantly affected by their own shocks, and these effects are anticipated well in advance, while their unexpected effects (for $j > 0$) are negligible. Also in the corresponding causal recursive VAR model, the variables are significantly affected by their own shocks, but, by construction, not anticipated. In view of the evidence in favor of the noncausal VAR, the latter model is misspecified, and thus the significant unexpected impact of the shocks that it implies, is misleading.

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

greater than 100, giving decisive evidence against Granger causality.

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. 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 fit and out-of-sample forecasting 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 theory, 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, or both. Moreover, according to our impulse response analysis, the economic shocks are highly anticipated, which undermines the validity of impulse response analysis based on the corresponding causal VAR model.

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 the seven variables included in the US macroeconomic model of Smets and 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.

ACKNOWLEDGEMENTS

We would like to thank Fabio Canova (co-editor) and the three anonymous referees for useful comments. 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.

References

Alessi L, Barigozzi M, Capasso M. 2008. Non-fundamentalness in structural econometric models: A review. *International Statistical Review* **79**: 16–47.

Bauwens L, Koop G, Korobilis D, Rombouts JVK. 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, Douc R, Guillin A, Marin JM, Robert CP. 2008. Adaptive importance sampling in general mixture classes. *Statistics and Computing* **18**: 447–459.

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

Davis RA, Song L. 2012. Noncausal Vector AR processes with application to economic time series. Working Paper. Columbia University.

Del Negro M, Schorfheide F. 2011. Bayesian macroeconometrics. In *Oxford Handbook of Bayesian Econometrics*, Geweke J, Koop G, van Dijk H (eds). Oxford University Press: Oxford; 293–389.

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

Gelfand A, Dey D. 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. 2005. *Contemporary Bayesian Econometrics and Statistics.* Wiley: New Jersey.

Gourieroux C, Jasiak J. 2014. Filtering and Prediction in noncausal processes. Centre de Recherche en Economie et Statistique, Working Papers 2014-15.

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

Hoogerheide, L, Opschoor A, van Dijk HK. 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 KR, Karlsson S. 1997. Numerical Methods for Estimation and Inference in Bayesian VAR-Models. *Journal of Applied Econometrics* **12**: 99-132.

Kohn R. 1979. Asymptotic estimation and hypothesis testing results for vector linear time series models. *Econometrica* **47**: 1005–1029.

Lanne M, Luoma A, Luoto J. 2012. Bayesian model selection and forecasting in noncausal autoregressive models. *Journal of Applied Econometrics* **27**: 812–830.

Lanne M, Luoto J. 2013. Does output gap, labour's share or unemployment rate drive inflation? *Oxford Bulletin of Economics and Statistics* **76**: 717–726.

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

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

Litterman RB. 1986. Forecasting with Bayesian vector autoregressions-five years of experience. *Journal of Business & Economic Statistics* **4**: 25–38.

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, Saikkonen P. 2013. Forecasting with a noncausal VAR model. *Computational Statistics and Data Analysis* **76**: 536–555.

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

Table I. Model selection.

Model	ln ML	Std.err.
VAR(2,0)	–936.48	0.0035
VAR(1,1)	–934.61	0.0066
VAR(0,2)	–935.45	0.0055
VAR(3,0)	–938.51	0.0045
VAR(2,1)	–935.16	0.0071
VAR(1,2)	–932.15	0.0045
VAR(0,3)	–934.95	0.0056
VAR(4,0)	–942.49	0.0089
VAR(3,1)	–938.54	0.0280
VAR(2,2)	–935.31	0.0192
VAR(1,3)	–934.84	0.0090
VAR(0,4)	–938.17	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 II. 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)	–701.8	–757.8	–796.1

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 III. 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)	1.150	1.513	1.980	–253.5	–304.7	–349.3
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)	3.448	3.617	3.563	–449.0	–453.9	–445.3
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.