<https://helda.helsinki.fi>

Bayesian Model Selection and Forecasting in Noncausal Autoregressive Models

Lanne, Markku

2012

Lanne , M , Luoma , A & Luoto , J 2012 , ' Bayesian Model Selection and Forecasting in Noncausal Autoregressive Models ' , Journal of Applied Econometrics , vol. 27 , no. 5 , pp. 812-830 . <https://doi.org/10.1002/jae.1217>

<http://hdl.handle.net/10138/230225>

<https://doi.org/10.1002/jae.1217>

acceptedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Bayesian Model Selection and Forecasting in Noncausal Autoregressive Models

Markku Lanne

Department of Political and Economic Studies and HECER

University of Helsinki

FINLAND

Arto Luoma

Department of Mathematics and Statistics

University of Tampere

FINLAND

Jani Luoto*

Department of Political and Economic Studies and HECER

University of Helsinki

FINLAND

Summary

In this paper, we propose a Bayesian estimation and forecasting procedure for noncausal autoregressive (AR) models. Specifically, we derive the joint posterior density of the past and future errors and the parameters, yielding predictive densities as a by-product. We show that the posterior model probabilities provide a convenient model selection criterion in discriminating between alternative causal and noncausal specifications. As an empirical application, we consider U.S. inflation. The posterior probability of noncausality is found to be high, over 98 percent. Furthermore, the purely noncausal specifications yield more accurate inflation forecasts than alternative causal and noncausal AR models.

JEL Classification: C11, C32, C52, E31

Keywords: Noncausality, Autoregression, Bayesian model selection, Forecasting

* Corresponding author: Jani Luoto, P.O. Box 17 (Arkadiankatu 7), 00014 University of Helsinki, Finland, Phone +358-(0)9-1911 28718, Fax +358-(0)9-191 28736, Email jani.luoto@helsinki.fi.

1. Introduction

Univariate autoregressive (AR) models have several uses in analyzing economic time series. However, it is the causal AR model that has almost exclusively been employed in econometrics although noncausal models have also to some extent been considered in statistics in general.¹ The main difference between the causal and noncausal AR models is that the latter allow for dependence on future as well as past values of the variable in question, whereas the former force the variable to depend only on its past. In the areas of economics where AR models are employed, expectations typically play a central role, and, therefore, extensions to noncausal models are likely to open up new possibilities, because they make explicit the dependence on future errors and values of the variable in question.

As discussed by Lanne and Saikkonen (2008), noncausality may be generated in different ways. In one case, already pointed out by Hansen and Sargent (1991), the noncausal autoregressive model arises as a nonfundamental solution of a rational expectations model, when the agents' information set is greater than that of the econometrician, who is estimating only a univariate model. The emergence of noncausality is an indication of the agents' being able to partly forecast future values of the economic variable in question by information unknown to the econometrician. Besides this case related to the discrepancy between the econometrician's and agents' information sets, heterogeneous beliefs have

¹ Calling these models (non)causal *filters* instead of autoregressions might be more appropriate, as pointed out by a referee. However, we have chosen to stick to the terminology established in the previous literature (see, e.g., Breidt *et al.*, 1991; Brockwell and Davis, 1991).

been shown to be a potential cause of nonfundamental solutions with nonrevealing equilibria, giving rise to a noncausal model (see, e.g., Kasa *et al.*, 2007).

The literature on noncausal AR models is not voluminous, and so far very few economic applications exist. Apart from Lanne and Saikkonen (2008), who found strong support for a noncausal AR specification for the U.S. inflation, previous studies on noncausal AR and related models in statistics only contain brief illustrations of the methods using economic data, but no serious applications.² Lanne and Saikkonen (2008) recently introduced a new formulation of the noncausal AR model that has a number of statistical advantages in addition to allowing for a convenient interpretation in terms of expectations, likely to be useful in economic applications. They also derived an approximate maximum likelihood estimator of this formulation and the related asymptotic distribution theory. Because causality and noncausality are not distinguishable under Gaussian errors, Lanne and Saikkonen (2008) suggested using Student's t -distribution, which seems appropriate in view of the fact that in economic applications residuals often turn out to deviate from normality in the direction of excess kurtosis. However, within their formulation of the model that we also consider in this paper, various alternative distributional assumptions can also be entertained.

Allowing for noncausality complicates model selection. Even if an economic variable can be assumed to be characterized as an AR process, causality or noncausality cannot be

² Breidt *et al.* (2001) fit a noncausal first-order AR model to a daily time series of the trading volume of Microsoft stock and the closely related all-pass model to the New Zealand/U.S. exchange rate. Huang and Pawitan (2000) applied a noninvertible moving average model to the U.S. unemployment rate.

determined on the basis of its autocorrelation structure because there are multiple causal and noncausal models of the same order producing identical autocorrelation functions. As pointed out above, with Gaussian errors, alternative causal and noncausal models of the same order also produce the same value of the likelihood function. Therefore, a non-Gaussian error distribution must be assumed, but even in that case model selection cannot be based on testing in a straightforward way because the alternative specifications are not nested. Following Breidt *et al.* (1991), Lanne and Saikkonen (2008) proposed a model selection procedure based on the maximum value of the likelihood function over a number of different model specifications of the same order and subsequent diagnostic checks for the adequacy of the model proposed by this criterion. In this paper, we consider the Bayesian analysis of noncausal AR models. We adopt the formulation of Lanne and Saikkonen (2008) and concentrate on model selection, since the nonnestedness of the models to be compared poses no particular problem in Bayesian analysis. Specifically, we use the posterior model probability of a particular specification to assess the degree of support in the data for that specification. The posterior model probabilities are based on an arbitrarily accurate approximation of the exact likelihood function where the past and future errors are considered unknown parameters. Thus, the inference is not conditional on initial values.

Our simulation experiments indicate that the proposed Bayesian model selection criterion works well in discriminating between causal and noncausal AR models. In particular, the expected posterior probabilities in favor of noncausal processes are high even under a relatively low degree of noncausality. On the other hand, when the true data generating

process is causal, our criterion selects the noncausal model markedly less frequently than the procedure of Breidt *et al.* (1991) and Lanne and Saikkonen (2008).

In addition to model selection, forecasting with the noncausal AR model is also considered. In particular, for optimal prediction of a noncausal process, knowledge of future errors is required. Because our approach treats these as unknown parameters, it has the advantage of providing a straightforward way to generate forecasts. As a matter of fact, to the best of our knowledge, this study is the first to provide a method of computing forecasts from these models.

We consider an empirical application to the same U.S. inflation series that Lanne and Saikkonen (2008) used, and in accordance with their results, we find support for the purely noncausal AR model, where current inflation only depends on expected future inflation. However, the posterior probability of the purely noncausal process is only about 30%. Taken at face value, this finding indicates that the observed persistence in inflation is caused partly by the predictability of inflation and partly by price-setting being dependent on past inflation. This goes contrary to typical New Keynesian models with forward-looking dynamics, but accords with much of the recent empirical literature. In contrast to that literature relying on causal AR models, though, we find expectations of future inflation to be the most important factor causing persistence. From the viewpoint of economics, this kind of availability of a measure of the likelihood of the purely noncausal model vis-à-vis the alternative AR models is probably the greatest value-added of Bayesian over classical analysis. We also conducted an out-of-sample forecasting exercise. The forecasts based on purely noncausal AR models turned out to be markedly

more accurate than those based on alternative AR models, especially at longer forecast horizons. These findings lend further support to future expectations playing a dominant role in the inflation process.

The plan of the paper is as follows. In Section 2, the noncausal AR model is presented and the likelihood function is derived. In Section 3, the choice of prior distributions is discussed. Section 4 shows how posterior analysis can be conducted. In Section 5, we describe the principles of model selection and present the results of the related simulation study. Section 6 presents the empirical application to the U.S. inflation. Finally, Section 7 concludes.

2. The model

Consider a stochastic process, y_t ($t = 0, \pm 1, \pm 2, \dots$), generated by

$$\phi(L^{-1})\rho(L)y_t = \varepsilon_t, \quad (1)$$

where $\phi(L^{-1}) = 1 - \phi_1 L^{-1} - \dots - \phi_s L^{-s}$, $\rho(L) = 1 - \rho_1 L - \dots - \rho_r L^r$, ε_t is a sequence of i.i.d. random variables with zero mean and scale parameter σ , and L is the lag operator. The polynomials $\phi(z)$ and $\rho(z)$ are assumed to have their zeros outside the unit circle so that

$$\rho(z) \neq 0 \text{ for } |z| \leq 1 \quad \text{and} \quad \phi(z) \neq 0 \text{ for } |z| \leq 1. \quad (2)$$

We use the abbreviation $AR(r,s)$ for the model defined by (1) and sometimes write $AR(r)$ for $AR(r, 0)$. The conventional causal $AR(r)$ model in which y_t depends on its past but not future values is obtained from (1) by setting $\phi_1 = \dots = \phi_s = 0$. The more interesting cases from the viewpoint of this paper arise, when this restriction does not hold. The process is said to be purely noncausal with dependence on future values only when $\rho_1 = \dots = \rho_r = 0$. In the mixed $AR(r,s)$ model where neither restriction holds, y_t depends on its past as well as future values.

Under the former condition in (2) the process $u_t = \phi(L^{-1})y_t$ has the following moving average representation,

$$u_t = \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j}, \quad (3)$$

with α_j the coefficient of z^j in the power series expansion of $\rho(L)^{-1}$. Similarly, by the latter condition in (2), the process $v_t = \rho(L)y_t$ can be written as

$$v_t = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t+j}, \quad (4)$$

with β_j the coefficient of z^j in the power series expansion of $\phi(L^{-1})^{-1}$. In sum, equation (1) can be expressed either in terms of u_t or v_t , that is, $\rho(L)u_t = \phi(L^{-1})v_t = \varepsilon_t$, and given the stationarity assumption in (2), the coefficients α_j and β_j decay to zero as $j \rightarrow \infty$.

For the Bayesian analysis of the noncausal model specified in equation (1), we need to derive the joint probability of the observations conditional on the parameters, i.e., the likelihood function, and specify the prior distributions of the parameters. Let us start with the likelihood function and defer the priors to Section 3.

Lanne and Saikkonen (2008) studied the maximum likelihood (ML) estimation of the noncausal model (1). Specifically, they showed that the vectors $\mathbf{z} = (u_1, \dots, u_r, \varepsilon_{r+1}, \dots, \varepsilon_{T-s}, v_{T-s+1}, \dots, v_T)$ and $\mathbf{y} = (y_1, y_2, \dots, y_T)$ are related by a linear transformation $\mathbf{z}' = \mathbf{B}\mathbf{A}\mathbf{y}'$ where the matrices $\mathbf{B}$ and $\mathbf{A}$ depend on $\boldsymbol{\rho} = (\rho_1, \dots, \rho_r)'$ and $\boldsymbol{\phi} = (\phi_1, \dots, \phi_s)'$. The determinant of $\mathbf{B}$ is unity, whereas $|\mathbf{A}|$ is the Jacobian determinant of the linear transformation given by equations $u_1 = \phi(L^{-1})y_1, u_2 = \phi(L^{-1})y_2, \dots, u_{T-s} = \phi(L^{-1})y_{T-s}, v_{T-s+1} = \rho(L)y_{T-s+1}, \dots, v_T = \rho(L)y_T$. Thus, given that the components $(u_1, \dots, u_r)$, $(\varepsilon_{r+1}, \dots, \varepsilon_{T-s})$ and $(v_{T-s+1}, \dots, v_T)$ of $\mathbf{z}$ are independent, as can be seen from equations (3) and (4), the exact likelihood can be expressed as

$$p(\mathbf{y}|\boldsymbol{\theta}) = p(\phi(L^{-1})y_1, \dots, \phi(L^{-1})y_r|\boldsymbol{\theta})p(\rho(L)y_{T-s+1}, \dots, \rho(L)y_T|\boldsymbol{\theta}) \\ \times |\mathbf{A}(\boldsymbol{\phi}, \boldsymbol{\rho})| \prod_{t=r+1}^{T-s} f_{\sigma}(\phi(L^{-1})\rho(L)y_t; \mathbf{v}), \quad (5)$$

where $\mathbf{v}$ is an additional parameter vector consisting of the parameters that determine the shape of the error distribution, and $\boldsymbol{\theta} = (\boldsymbol{\rho}', \boldsymbol{\phi}', \sigma, \mathbf{v})'$ is a vector of all parameters, and the notation $\mathbf{A}(\boldsymbol{\phi}, \boldsymbol{\rho})$ is used to indicate the dependence of $\mathbf{A}$ on $\boldsymbol{\rho}$ and $\boldsymbol{\phi}$. The error distribution is assumed to be non-Gaussian with density $f_{\sigma}(x) = \sigma^{-1}f(\sigma^{-1}x; \mathbf{v})$. As in Lanne and Saikkonen (2008), the density function $f_{\sigma}(\cdot)$ satisfies the regularity conditions of Andrews

et al. (2006) which, among other things, require that $f_\sigma(\cdot)$ is twice continuously differentiable with respect to x and $\mathbf{v}$, and positive for all real numbers x and all permissible values of the $\mathbf{v}$.

Following Lanne and Saikkonen (2008), we assume that ε_t follows Student's t distribution with scale parameter $\omega^{-1/2} = \sigma$ and ν degrees of freedom. To the best of our knowledge, the densities $p(\phi(L^{-1})y_1, \dots, \phi(L^{-1})y_r | \boldsymbol{\theta})$ and $p(\rho(L)y_{T-s+1}, \dots, \rho(L)y_T | \boldsymbol{\theta})$ cannot be presented in closed form in this case. With the errors ε_t being independently t -distributed, a natural suggestion would be a multivariate t distribution. However, it is easy to see that this cannot be the case. It is well-known that the joint distribution of independently t -distributed random variables is not a multivariate t -distribution, and as Tarami and Pourahmadi (2000), *inter alia*, point out, the joint distribution of linear transformations of ε_t such as u_t and v_t , is a multivariate t -distribution only if ε_t 's follow a multivariate t -distribution.

Therefore, Lanne and Saikkonen (2008) approximate the likelihood function using the last factor on the right hand side of (5). We consider an alternative approach which employs the approximations $u_t \approx \sum_{j=0}^M \alpha_j \varepsilon_{t-j}$ and $v_t \approx \sum_{j=0}^M \beta_j \varepsilon_{t+j}$, where the integer M is assumed to be sufficiently large to make the approximation error negligible. Our idea is straightforward. We are not able to express the joint distribution of $u_1, \dots, u_r$ and $v_{T-s+1}, \dots, v_T$ in closed form, but we can write their conditional densities, given the errors $\boldsymbol{\varepsilon}^- = (\varepsilon_M, \dots, \varepsilon_{-1}, \varepsilon_0)$ and $\boldsymbol{\varepsilon}^+ = (\varepsilon_{T+1}, \varepsilon_{T+2}, \dots, \varepsilon_{T+M})$, in terms of $\varepsilon_1, \dots, \varepsilon_r$ and $\varepsilon_{T-s+1}, \dots, \varepsilon_T$ using the change of variable theorem. In practice, an arbitrarily accurate approximation of the

likelihood function (5) can be obtained by starting with the joint density $p(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta})$ and then integrating out the past and future errors:

$$p(\mathbf{y} | \boldsymbol{\theta}) \approx \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta}) d\boldsymbol{\varepsilon}^- d\boldsymbol{\varepsilon}^+. \quad (6)$$

In general, this integral cannot be computed analytically. However, an applicable numerical solution can be obtained once the distribution of the errors ε_t has been chosen. In the following, we shall describe how such a solution can be obtained.

According to equations (3) and (4), the blocks $(\boldsymbol{\varepsilon}^-, u_1, \dots, u_r)$ and $(v_{T-s+1}, \dots, v_T, \boldsymbol{\varepsilon}^+)$ are independent of each other and independent of $\varepsilon_{r+1}, \dots, \varepsilon_{T-s}$. The joint density $p(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta})$ can therefore be expressed as (cf. (5))

$$\begin{aligned} p(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta}) &= p(\boldsymbol{\varepsilon}^-, \phi(L^{-1})y_1, \dots, \phi(L^{-1})y_r | \boldsymbol{\theta}) \\ &\quad \times |\mathbf{A}(\boldsymbol{\phi}, \boldsymbol{\rho})| \prod_{t=r+1}^{T-s} f_{\sigma}(\phi(L^{-1})\rho(L)y_t; \mathbf{v}) \\ &\quad \times p(\rho(L)y_{T-s+1}, \dots, \rho(L)y_T, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta}). \end{aligned} \quad (7)$$

Using the basic rules of conditional probability, this can be rewritten as

$$\begin{aligned} p(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta}) &= \prod_{t=-M}^0 f_{\sigma}(\varepsilon_t; \mathbf{v}) p(\phi(L^{-1})y_1, \dots, \phi(L^{-1})y_r | \boldsymbol{\theta}, \boldsymbol{\varepsilon}^-) \\ &\quad \times |\mathbf{A}(\boldsymbol{\phi}, \boldsymbol{\rho})| \prod_{t=r+1}^{T-s} f_{\sigma}(\phi(L^{-1})\rho(L)y_t; \mathbf{v}) \end{aligned}$$

$$\times p(\rho(L)y_{T-s+1}, \dots, \rho(L)y_T | \boldsymbol{\theta}, \boldsymbol{\varepsilon}^+) \prod_{t=T+1}^{T+M} f_\sigma(\varepsilon_t; \mathbf{v}). \quad (8)$$

Recalling that $u_t = \phi(L^{-1})y_t$, $v_t = \rho(L)y_t$ and $\rho(L)u_t = \phi(L^{-1})v_t = \varepsilon_t$, and noticing that the Jacobian determinants of the transformations $(u_1, \dots, u_r) \rightarrow (\varepsilon_1, \dots, \varepsilon_r)$ and $(v_{T-s+1}, \dots, v_T) \rightarrow (\varepsilon_{T-s+1}, \dots, \varepsilon_T)$ are unity, equation (8) can in turn be written as

$$\begin{aligned} p(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta}) &= \prod_{t=-M}^0 f_\sigma(\varepsilon_t; \mathbf{v}) \prod_{t=1}^r f_\sigma(\rho(L)u_t(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^-, \mathbf{y}); \mathbf{v}) \\ &\quad \times |\mathbf{A}(\boldsymbol{\phi}, \boldsymbol{\rho})| \prod_{t=r+1}^{T-s} f_\sigma(\phi(L^{-1})\rho(L)y_t; \mathbf{v}) \\ &\quad \times \prod_{t=T-s+1}^T f_\sigma(\phi(L^{-1})v_t(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \mathbf{y}); \mathbf{v}) \prod_{t=T+1}^{T+M} f_\sigma(\varepsilon_t; \mathbf{v}). \end{aligned} \quad (9)$$

We have used the notation $v_t(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \mathbf{y})$ to emphasize that, although $v_{T-s+1}, \dots, v_T$ can be calculated directly from the data using $\boldsymbol{\rho}$, the post-sample values $v_{T+1}, \dots, v_{T+s}$ depend on $\boldsymbol{\varepsilon}^+$ and $\boldsymbol{\phi}$. Similarly, given $\boldsymbol{\phi}$, $u_1, \dots, u_r$ can be calculated from the data, whereas the pre-sample values $u_{1-r}, \dots, u_0$ depend on $\boldsymbol{\varepsilon}^-$ and $\boldsymbol{\rho}$ (hence the notation $u_t(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^-, \mathbf{y})$). In other words, conditional on $\boldsymbol{\varepsilon}^-$ and $\boldsymbol{\varepsilon}^+$, the values $u_{1-r}, \dots, u_0$ and $v_{T+1}, \dots, v_{T+s}$ can be readily calculated using the approximations $u_t \approx \sum_{j=0}^M \alpha_j \varepsilon_{t-j}$ and $v_t \approx \sum_{j=0}^M \beta_j \varepsilon_{t+j}$, respectively, whereas given $u_{1-r}, \dots, u_0$ and $v_{T+1}, \dots, v_{T+s}$, the transformations $(u_1, \dots, u_r) \rightarrow (\varepsilon_1, \dots, \varepsilon_r)$ and $(v_{T-s+1}, \dots, v_T) \rightarrow (\varepsilon_{T-s+1}, \dots, \varepsilon_T)$ are standard. The accuracy of the approximation in (6) therefore depends on the coefficients α_j and β_j . In practice, M should be set so large that the coefficients α_j and β_j are practically zero for $j > M$, resulting in a negligible approximation error. The appropriate value of M depends on the zeros of the polynomials

$\rho(z)$ and $\phi(z)$; the closer they are to the unit circle the larger the value of M that should be used. Notice that, in extreme cases when the zeros of the polynomials $\rho(z)$ and $\phi(z)$ are close to the unit circle causing the approximation (6) to become inaccurate with a decent M , the following alternative calculations of $v_{T+1}, \dots, v_{T+s}$ and $u_{1-r}, \dots, u_0$ seem to work well in practice (otherwise these two ways produce virtually the same post-sample and pre-sample values). From equation (1) we have

$$\begin{aligned}
 v_{T+M} &= \varepsilon_{T+M} + \phi_1 v_{T+M+1} + \dots + \phi_s v_{T+M+s}, \\
 v_{T+M-1} &= \varepsilon_{T+M-1} + \phi_1 v_{T+M} + \dots + \phi_s v_{T+M+s-1}, \\
 &\vdots \\
 v_{T+1} &= \varepsilon_{T+1} + \phi_1 v_{T+2} + \dots + \phi_s v_{T+s+1},
 \end{aligned} \tag{10}$$

and plugging in the simulated values of $\varepsilon_{T+1}, \dots, \varepsilon_{T+M}$ and setting $v_{T+M+1}, \dots, v_{T+M+s}$ at their expected value 0, we get $v_{T+1}, \dots, v_{T+M}$. Similarly,

$$\begin{aligned}
 u_{-M} &= \varepsilon_{-M} + \rho_1 u_{-M-1} + \dots + \rho_r u_{-M-r}, \\
 u_{-M+1} &= \varepsilon_{-M+1} + \rho_1 u_{-M} + \dots + \rho_r u_{-M-r+1}, \\
 &\vdots \\
 u_0 &= \varepsilon_0 + \rho_1 u_{-1} + \dots + \rho_r u_{-r},
 \end{aligned} \tag{11}$$

where $u_{-M-1}, \dots, u_{-M-r}$, in turn, are set at zero.

In what follows, we assume that the elements of $\boldsymbol{\varepsilon}^-$ and $\boldsymbol{\varepsilon}^+$ are unobserved variables, whose posterior densities are obtained by simulation methods along with the unknown parameters. This, of course facilitates a numerical solution to the integral in (6). Furthermore, as pointed out above, explicit incorporation of the past and future errors into the analysis allows for computing optimal forecasts from the noncausal model (1) in a straightforward manner. Specifically, the posterior density of v_{T+h} ($h > 0$) may be simulated using the approximation $v_{T+h} \approx \sum_{j=0}^{M-h} \beta_j \varepsilon_{T+h+j}$, given the posterior simulations of $\boldsymbol{\varphi}$ and $\varepsilon_{T+h}, \varepsilon_{T+2}, \dots, \varepsilon_{T+M}$. If $r > 0$, the predictive density of y_{T+h} can be calculated using the recursive formula $y_{T+h} = \rho_1 y_{T+h-1} + \dots + \rho_r y_{T+h-r} + v_{T+h}$. The means or medians of these predictive distributions can be used as point forecasts. When the process is purely noncausal (that is, $r = 0$), the predictive densities of v_{T+h} and y_{T+h} coincide.

3. Priors

In addition to the likelihood function derived in Section 2, Bayesian analysis requires the specification of prior distributions of the parameters of interest. We assume the prior independence of the parameters, which in the case of Student's t error distribution comprise $\boldsymbol{\rho}$, $\boldsymbol{\varphi}$, σ and v . Under this assumption their joint prior distribution can thus be expressed as

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\rho})p(\boldsymbol{\varphi})p(\sigma)p(v). \quad (12)$$

We use proper priors for $\boldsymbol{\rho}$ and $\boldsymbol{\varphi}$ because when improper priors, i.e. priors that are not well-defined density functions, are used for parameters occurring in one model but not

the other, posterior odds ratios are not identified (see, e.g., Bartlett, 1957; O'Hagan, 1995). In particular, we shall adopt the following multivariate normal prior distributions, truncated to the stationary region to ensure that the assumptions given in (2) hold,

$$p(\boldsymbol{\rho}) \propto N(\boldsymbol{\rho} | \boldsymbol{\rho}_0, \mathbf{P}_0^{-1}) \times I_S(\boldsymbol{\rho}); \quad p(\boldsymbol{\varphi}) \propto N(\boldsymbol{\varphi} | \boldsymbol{\varphi}_0, \boldsymbol{\Phi}_0^{-1}) \times I_S(\boldsymbol{\varphi}), \quad (13)$$

where $I_S(\boldsymbol{\rho})$ and $I_S(\boldsymbol{\varphi})$ are indicator functions equaling unity in the stationary region and zero otherwise, and $\boldsymbol{\rho}_0$, $\boldsymbol{\varphi}_0$, $\mathbf{P}_0$ and $\boldsymbol{\Phi}_0$ are known hyperparameters. We further assume that $\boldsymbol{\rho}_0$ and $\boldsymbol{\varphi}_0$ are zero vectors. The prior precision matrices $\mathbf{P}_0$ and $\boldsymbol{\Phi}_0$ are parameterized according to the Minnesota tradition by assuming that the probability of a coefficient being close to zero increases with the lag (and lead) (see Litterman, 1986). In practice, $\mathbf{P}_0$ and $\boldsymbol{\Phi}_0$ are assumed to be diagonal, and the prior standard deviations of ϕ_i and ρ_j are set at k/i , $i = 1, \dots, r$, and k/j , $j = 1, \dots, s$, respectively. However, because we do not have much experience of how this shrinkage prior works in forecasting with noncausal models, we shall also check the forecast accuracy of purely noncausal models when $\boldsymbol{\Phi}_0$ is an identity matrix. The scalar k is set at unity because when posterior odds are used in model comparison, a value of k substantially less than 1 (uninformative prior) typically penalizes long lags and leads, whereas a value of k greater than 1 (informative prior) favors long lags and leads. Also notice that only the kernels of the priors in (13) are known when the number of lags or leads is greater than unity. The integrating constants of these priors are estimated by drawing from $N(\boldsymbol{\rho} | \boldsymbol{\rho}_0, \mathbf{P}_0^{-1})$ or $N(\boldsymbol{\varphi} | \boldsymbol{\varphi}_0, \boldsymbol{\Phi}_0^{-1})$, and calculating the proportion of draws within the stationary region (see, e.g., Koop, 2003). These estimates are based on 1000,000 draws.

For $\omega = \sigma^{-2}$ the standard noninformative prior $p(\omega) \propto 1/\omega$ is selected. An improper prior can be used, since this parameter is common to all models, implying that its arbitrary integrating constant does not affect the posterior odds ratios. Finally, as shown by Bauwens and Lubrano (1998), sufficient prior information is needed on the degrees of freedom parameter ν in Student's t distribution to force the posterior, in order to be integrable, to tend to zero quickly enough at the tail. We will follow Geweke (1993) in using an exponential density. In particular, $p(\nu) = (1/\nu_0)\exp(-(2-\nu)/\nu_0)$ when $\nu > 2$. In our simulation study, we set ν_0 at 3. In the empirical application we give more weight to the data and set ν_0 at 5.

4. Posterior analysis

With the joint density of $(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+)$ derived in Section 2 and the prior distribution specified in Section 3, we are able to compute the posterior distribution of the parameters $\boldsymbol{\theta}$ and the past and future unobserved errors. Let $p(\mathbf{y})$ denote the joint density of the data. The joint posterior density of $\boldsymbol{\varepsilon}^-, \boldsymbol{\varepsilon}^+$ and $\boldsymbol{\theta}$ can then be expressed as

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\varepsilon}^-, \boldsymbol{\varepsilon}^+ | \mathbf{y}) &= \frac{p(\boldsymbol{\theta})p(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta})}{p(\mathbf{y})} \\ &= \frac{p(\boldsymbol{\theta})p(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta})}{\int \cdots \int p(\boldsymbol{\theta})p(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta}) d\boldsymbol{\varepsilon}^- d\boldsymbol{\varepsilon}^+ d\boldsymbol{\theta}}. \end{aligned} \quad (14)$$

It is obvious that a closed form solution exists for neither the marginal likelihood $p(\mathbf{y})$ nor the posterior moments of the parameters, and numerical methods are required. To build a

decent posterior sampler with a high acceptance rate of candidate draws, under the assumption that in (1) each disturbance ε_t follows Student's t distribution with location parameter 0, scale $\omega^{-1/2}$, and ν degrees of freedom, that is, $\varepsilon_t \sim t(0, \omega^{-1}; \nu)$, we employ the following representation

$$\varepsilon_t = \tilde{\omega}_t^{-1/2} \eta_t, \quad (t = 1, \dots, T) \quad (15)$$

where η_t are i.i.d. $N(0, \omega^{-1})$ and $\nu \tilde{\omega}_t$ are i.i.d. $\chi^2(\nu)$ (chi-square with ν degrees of freedom), (see West, 1984; Geweke, 1993, 2005). Using the above representation, equation (9) can be rewritten as

$$\begin{aligned} p(\boldsymbol{\varepsilon}^+, \mathbf{y}, \boldsymbol{\varepsilon}^- | \boldsymbol{\theta}) &\propto \prod_{t=-M}^0 f_\sigma(\varepsilon_t; \nu) \omega^{r/2} \prod_{t=1}^r \tilde{\omega}_t^{1/2} \exp \left[-\frac{\omega}{2} \sum_{t=1}^r \tilde{\omega}_t [\rho(L) u_t(\boldsymbol{\varphi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^-, \mathbf{y})]^2 \right] \\ &\quad \times |\mathbf{A}(\boldsymbol{\varphi}, \boldsymbol{\rho})| \omega^{(T-s-r)/2} \prod_{t=r+1}^{T-s} \tilde{\omega}_t^{1/2} \exp \left[-\frac{\omega}{2} \sum_{t=r+1}^{T-s} \tilde{\omega}_t (\phi(L^{-1}) \rho(L) y_t)^2 \right] \\ &\quad \times \omega^{s/2} \prod_{t=T-s+1}^T \tilde{\omega}_t^{1/2} \exp \left[-\frac{\omega}{2} \sum_{t=T-s+1}^T \tilde{\omega}_t [\phi(L^{-1}) v_t(\boldsymbol{\varphi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \mathbf{y})]^2 \right] \prod_{t=T+1}^{T+M} f_\sigma(\varepsilon_t; \nu). \quad (16) \end{aligned}$$

Recalling that $\rho(L) \phi(L^{-1}) y_t = \rho(L) u_t = \phi(L^{-1}) v_t = \varepsilon_t$, this can in turn be expressed as

$$\begin{aligned} p(\boldsymbol{\varepsilon}^+, \mathbf{y}, \boldsymbol{\varepsilon}^- | \boldsymbol{\theta}) &\propto |\mathbf{A}(\boldsymbol{\varphi}, \boldsymbol{\rho})| \prod_{t=-M}^0 t(\varepsilon_t | 0, \omega^{-1}; \nu) \prod_{t=T+1}^{T+M} t(\varepsilon_t | 0, \omega^{-1}; \nu) q_1(\boldsymbol{\varphi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \mathbf{y}) \\ &\quad \times \omega^{(T-s)/2} |\boldsymbol{\Omega}_u|^{1/2} \exp \left[-\frac{\omega}{2} (\mathbf{u} - \mathbf{U}\boldsymbol{\rho})' \boldsymbol{\Omega}_u (\mathbf{u} - \mathbf{U}\boldsymbol{\rho}) \right] \end{aligned}$$

$$\begin{aligned}
&= |\mathbf{A}(\boldsymbol{\phi}, \boldsymbol{\rho})| \prod_{t=-M}^0 t(\varepsilon_t | 0, \omega^{-1}; \nu) \prod_{t=T+1}^{T+M} t(\varepsilon_t | 0, \omega^{-1}; \nu) q_2(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^-, \mathbf{y}) \\
&\quad \times \omega^{(T-r)/2} |\boldsymbol{\Omega}_v|^{1/2} \exp \left[-\frac{\omega}{2} (\mathbf{v} - \mathbf{V}\boldsymbol{\phi})' \boldsymbol{\Omega}_v (\mathbf{v} - \mathbf{V}\boldsymbol{\phi}) \right], \quad (17)
\end{aligned}$$

where $\boldsymbol{\Omega}_u = \text{diag}(\tilde{\omega}_1, \dots, \tilde{\omega}_{T-s})$, $\boldsymbol{\Omega}_v = \text{diag}(\tilde{\omega}_{r+1}, \dots, \tilde{\omega}_T)$, $\mathbf{u} = (u_1, \dots, u_{T-s})'$, $\mathbf{v} = (v_{r+1}, \dots, v_T)'$, the t th rows of matrices $\mathbf{U}$ and $\mathbf{V}$ are given by $u_{t-1}, \dots, u_{t-r}$ and $v_{t+1}, \dots, v_{t+s}$, respectively, and $q_1(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \mathbf{y})$ and $q_2(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^-, \mathbf{y})$ are functions defined by

$$q_1(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \mathbf{y}) = \omega^{s/2} \prod_{t=T-s+1}^T \tilde{\omega}_t^{1/2} \exp \left[-\frac{\omega}{2} \sum_{t=T-s+1}^T \tilde{\omega}_t [\phi(L^{-1})v_t(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \mathbf{y})]^2 \right], \quad (18)$$

and

$$q_2(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^-, \mathbf{y}) = \omega^{r/2} \prod_{t=1}^r \tilde{\omega}_t^{1/2} \exp \left[-\frac{\omega}{2} \sum_{t=1}^r \tilde{\omega}_t [\rho(L)u_t(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^-, \mathbf{y})]^2 \right]. \quad (19)$$

Thus, conditional on $\boldsymbol{\Omega}_u$ and $\boldsymbol{\phi}$, or alternatively $\boldsymbol{\Omega}_v$ and $\boldsymbol{\rho}$, the last factor on the right hand side of (17) has the same form as the likelihood of the linear regression model. Given the marginal prior densities defined in Section 3, it is therefore straightforward to verify that the full conditional distributions for $\boldsymbol{\varepsilon}^+$, $\boldsymbol{\varepsilon}^-$, $\boldsymbol{\phi}$, $\boldsymbol{\rho}$, ω , $\tilde{\omega}_t$, and ν are given by

$$p(\boldsymbol{\varepsilon}^+ | \boldsymbol{\rho}, \boldsymbol{\phi}, \tilde{\boldsymbol{\omega}}, \nu, \omega, \mathbf{y}) \propto q_1(\boldsymbol{\varepsilon}^+) \prod_{t=T+1}^{T+M} t(\varepsilon_t | 0, \omega^{-1}; \nu), \quad (20)$$

$$p(\boldsymbol{\varepsilon}^- | \boldsymbol{\rho}, \boldsymbol{\phi}, \tilde{\boldsymbol{\omega}}, \nu, \omega, \mathbf{y}) \propto q_2(\boldsymbol{\varepsilon}^-) \prod_{t=-M}^0 t(\varepsilon_t | 0, \omega^{-1}; \nu), \quad (21)$$

$$p(\boldsymbol{\phi} | \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \boldsymbol{\varepsilon}^-, \tilde{\boldsymbol{\omega}}, \omega, \mathbf{y}) \propto q_2(\boldsymbol{\phi}) |\mathbf{A}(\boldsymbol{\phi})| N(\boldsymbol{\phi} | \bar{\boldsymbol{\phi}}, [\boldsymbol{\Phi}_0 + \omega \mathbf{V}' \boldsymbol{\Omega}_v \mathbf{V}]^{-1}) \times I_s(\boldsymbol{\phi}), \quad (22)$$

$$p(\boldsymbol{\rho}|\boldsymbol{\Phi}, \boldsymbol{\varepsilon}^+, \boldsymbol{\varepsilon}^-, \tilde{\boldsymbol{\omega}}, \omega, \mathbf{y}) \propto q_1(\boldsymbol{\rho})|\mathbf{A}(\boldsymbol{\rho})|N(\boldsymbol{\rho}|\bar{\boldsymbol{\rho}}, [\mathbf{P}_0 + \omega \mathbf{U}'\boldsymbol{\Omega}_u\mathbf{U}]^{-1}) \times I_S(\boldsymbol{\rho}), \quad (23)$$

$$p(\omega|\boldsymbol{\Phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \boldsymbol{\varepsilon}^-, \tilde{\boldsymbol{\omega}}, \nu, \mathbf{y}) \propto \prod_{t=-M}^0 t(\varepsilon_t|0, \omega^{-1}; \nu) \prod_{t=T+1}^{T+M} t(\varepsilon_t|0, \omega^{-1}; \nu) G\left(\omega \left| \frac{T}{2}, \frac{2}{S} \right.\right), \quad (24)$$

$$(\nu + \omega [\phi(L^{-1})\rho(L)y_t]^2) \tilde{\omega}_t | \boldsymbol{\Phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \boldsymbol{\varepsilon}^-, \omega, \nu, \mathbf{y} \sim \chi^2(\nu + 1), \quad (25)$$

and

$$\begin{aligned} p(\nu|\boldsymbol{\varepsilon}^+, \boldsymbol{\varepsilon}^-, \tilde{\boldsymbol{\omega}}, \omega, \mathbf{y}) &\propto \prod_{t=-M}^0 t(\varepsilon_t|0, \omega^{-1}; \nu) \prod_{t=T+1}^{T+M} t(\varepsilon_t|0, \omega^{-1}; \nu) \\ &\times [2^{\nu/2} \Gamma(\nu/2)]^{-T} \nu^{T\nu/2} \left(\prod_{t=1}^T \tilde{\omega}_t^{(\nu-2)/2} \right) \\ &\times \exp \left\{ - \left(\frac{1}{\nu_0} + \frac{1}{2} \sum_{t=1}^T \tilde{\omega}_t \right) \nu \right\}, \end{aligned} \quad (26)$$

where $\bar{\boldsymbol{\Phi}} = [\boldsymbol{\Phi}_0 + \omega \mathbf{V}'\boldsymbol{\Omega}_v\mathbf{V}]^{-1}[\boldsymbol{\Phi}_0\boldsymbol{\Phi}_0 + \omega \mathbf{V}'\boldsymbol{\Omega}_v\mathbf{V}]$, $\bar{\boldsymbol{\rho}} = [\mathbf{P}_0 + \omega \mathbf{U}'\boldsymbol{\Omega}_u\mathbf{U}]^{-1}[\mathbf{P}_0\boldsymbol{\rho}_0 + \omega \mathbf{U}'\boldsymbol{\Omega}_u\mathbf{u}]$, and $\tilde{\boldsymbol{\omega}} = (\tilde{\omega}_1, \dots, \tilde{\omega}_T)$, the $\tilde{\omega}_t$'s ($t = 1, \dots, T$) being conditionally independent, and $G(\omega|T/2, 2/S)$ is a Gamma density with shape $T/2$ and scale $2/S$, with $S = \sum_{t=1}^T \tilde{\omega}_t [\phi(L^{-1})\rho(L)y_t]^2$. Furthermore, for fixed values of $(\boldsymbol{\Phi}, \boldsymbol{\varepsilon}^+)$ and $(\boldsymbol{\rho}, \boldsymbol{\varepsilon}^-)$ the functions $q_1(\boldsymbol{\Phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \mathbf{y})$ and $q_2(\boldsymbol{\Phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^-, \mathbf{y})$ are denoted as $q_1(\boldsymbol{\rho})$ and $q_2(\boldsymbol{\Phi})$, respectively, whereas for fixed values of $\boldsymbol{\rho}$ and $\boldsymbol{\Phi}$ they are denoted as $q_1(\boldsymbol{\varepsilon}^+)$ and $q_2(\boldsymbol{\varepsilon}^-)$. The same logic is used for $|\mathbf{A}(\boldsymbol{\Phi}, \boldsymbol{\rho})|$, producing the notations $|\mathbf{A}(\boldsymbol{\Phi})|$ and $|\mathbf{A}(\boldsymbol{\rho})|$. The conditional densities in (20) and (21) follow directly from the fact that $\boldsymbol{\varepsilon}^+$ and $\boldsymbol{\varepsilon}^-$ appear in (17) only through the factors $t(\varepsilon_{T+1}|0, \omega^{-1}; \nu) \cdots t(\varepsilon_{T+M}|0, \omega^{-1}; \nu) \cdot q_1(\boldsymbol{\Phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^+, \mathbf{y})$ and $t(\varepsilon_{-M}|0, \omega^{-1}; \nu) \cdots t(\varepsilon_0|0, \omega^{-1}; \nu) \cdot q_2(\boldsymbol{\Phi}, \boldsymbol{\rho}, \boldsymbol{\varepsilon}^-, \mathbf{y})$, respectively. By combining the truncated multivariate normal priors given in (13) with the density in (17), equations (22) and (23) can be derived using standard calculations (see, e.g., Zellner, 1971; Geweke, 1993). Given $\boldsymbol{\varepsilon}^+$, $\boldsymbol{\varepsilon}^-$, $\boldsymbol{\Phi}$, $\boldsymbol{\rho}$, and the marginal

priors, equations (24)-(26) are close to the corresponding conditional densities in Geweke (1993).

Only the density in (25) is standard and can be readily used to simulate random numbers. Therefore, we adopt a Metropolis-Hastings algorithm described below. According to (20), it seems reasonable to start by drawing a proposal $\boldsymbol{\varepsilon}_*^+$ from the density $t(\varepsilon_{T+1}|0, \omega^{-1}; \nu) \cdots t(\varepsilon_{T+M}|0, \omega^{-1}; \nu)$. Then, the definition of the Metropolis-Hastings algorithm implies that $\boldsymbol{\varepsilon}_*^+$ is accepted with probability

$$\alpha_{\varepsilon^+} = \min \left\{ 1, \frac{q_1(\boldsymbol{\varepsilon}_*^+)}{q_1(\boldsymbol{\varepsilon}_{(i)}^+)} \right\}, \quad (27)$$

where $\boldsymbol{\varepsilon}_{(i)}^+$ ($i = 1, \dots, N$) is the current simulated realization of $\boldsymbol{\varepsilon}^+$. If the draw is accepted $\boldsymbol{\varepsilon}_{(i+1)}^+$ is set at $\boldsymbol{\varepsilon}_*^+$, otherwise at $\boldsymbol{\varepsilon}_{(i)}^+$. Similarly, the density $t(\varepsilon_{-M}|0, \omega^{-1}; \nu) \cdots t(\varepsilon_0|0, \omega^{-1}; \nu)$ can be used to draw $\boldsymbol{\varepsilon}_*^-$, which is accepted with probability

$$\alpha_{\varepsilon^-} = \min \left\{ 1, \frac{q_2(\boldsymbol{\varepsilon}_*^-)}{q_2(\boldsymbol{\varepsilon}_{(i)}^-)} \right\}. \quad (28)$$

If $\boldsymbol{\varepsilon}_*^-$ is rejected, $\boldsymbol{\varepsilon}_{(i+1)}^-$ is set at its current value. Next, for given $\boldsymbol{\varphi}^{(i)}$ and $\boldsymbol{\rho}^{(i)}$, we shall form vectors $\mathbf{u}$ and $\mathbf{v}$, and calculate matrices $\mathbf{U}$ and $\mathbf{V}$ using $\boldsymbol{\varepsilon}_{(i+1)}^-$ and $\boldsymbol{\varepsilon}_{(i+1)}^+$, respectively. Then, the truncated normal density $N(\boldsymbol{\varphi} | \bar{\boldsymbol{\varphi}}, [\boldsymbol{\Phi}_0 + \omega \mathbf{V}' \boldsymbol{\Omega}_v \mathbf{V}]^{-1}) \times I_S(\boldsymbol{\varphi})$ can be successfully employed as a candidate density for $\boldsymbol{\varphi}$. In this step the draw $\boldsymbol{\varphi}_*$ is accepted with probability

$$\alpha_{\boldsymbol{\phi}} = \min \left\{ 1, \frac{|\mathbf{A}(\boldsymbol{\phi}_*)| q_2(\boldsymbol{\phi}_*)}{|\mathbf{A}(\boldsymbol{\phi}_{(i)})| q_2(\boldsymbol{\phi}_{(i)})} \right\}. \quad (29)$$

Drawing from $N(\boldsymbol{\phi} | \bar{\boldsymbol{\phi}}, [\boldsymbol{\Phi}_0 + \omega \mathbf{V}' \boldsymbol{\Omega}_v \mathbf{V}]^{-1}) \times I_s(\boldsymbol{\phi})$ can be done by sampling from the untruncated variant until the draw lies in the stationary region. In a similar manner, $N(\boldsymbol{\rho} | \bar{\boldsymbol{\rho}}, [\mathbf{P}_0 + \omega \mathbf{U}' \boldsymbol{\Omega}_u \mathbf{U}]^{-1}) \times I_s(\boldsymbol{\rho})$ is a suitable candidate density for $\boldsymbol{\rho}$ and the candidate $\boldsymbol{\rho}^*$ is accepted with probability

$$\alpha_{\boldsymbol{\rho}} = \min \left\{ 1, \frac{|\mathbf{A}(\boldsymbol{\rho}_*)| q_1(\boldsymbol{\rho}_*)}{|\mathbf{A}(\boldsymbol{\rho}_{(i)})| q_1(\boldsymbol{\rho}_{(i)})} \right\}. \quad (30)$$

In the special case when either r or s is zero, the full conditional distribution of $\boldsymbol{\phi}$ and $\boldsymbol{\rho}$ is of known form, since then $|\mathbf{A}(\boldsymbol{\phi}, \boldsymbol{\rho})| = 1$.

The candidate ω^* is drawn from $G(\omega | T/2, 2/S)$ and accepted with probability

$$\alpha_{\omega} = \min \left\{ 1, \frac{\prod_{t=-M}^0 t(\varepsilon_t | 0, \omega_*^{-1}; \nu) \prod_{t=T+1}^{T+M} t(\varepsilon_t | 0, \omega_*^{-1}; \nu)}{\prod_{t=-M}^0 t(\varepsilon_t | 0, \omega_{(i)}^{-1}; \nu) \prod_{t=T+1}^{T+M} t(\varepsilon_t | 0, \omega_{(i)}^{-1}; \nu)} \right\}. \quad (31)$$

A univariate normal distribution, with mean at the mode of (26) and precision equal to the negative of the second derivative of the log posterior, evaluated at the mode, can

successfully be used as a candidate distribution for ν and the acceptance-reception probability can be calculated using equation (26) (see Geweke, 2005). Finally, the convergence of the sampler can be monitored using univariate and multivariate potential scale reduction factors (PSRF) (see Gelman and Rubin, 1992; Brooks and Gelman, 1998).³

5. Model comparison

So far, we have assumed the noncausal AR model specification known, which, of course, is virtually never the case. Instead, the orders r and s of the polynomials $\rho(L)$ and $\phi(L^{-1})$, respectively, must in practice be determined based on the data. The nonnestedness of the alternative $AR(r,s)$ models complicates classical model selection, but poses no particular problems in Bayesian analysis, as model selection can be based on marginal likelihoods of the alternative models.

For each noncausal AR model specification, M_{ij} ($i = 0, \dots, r_{max}, j = 0, \dots, s_{max}$), the marginal likelihood is simply the denominator of the joint posterior density (14),

$$p(\mathbf{y}|M_{ij}) = \int \cdots \int p(\boldsymbol{\theta}|M_{ij}) p(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+ | \boldsymbol{\theta}, M_{ij}) d\boldsymbol{\varepsilon}^- d\boldsymbol{\varepsilon}^+ d\boldsymbol{\theta} . \quad (32)$$

³ In our study, chains of length 100,000 were simulated (using different starting values), and the first 10,000 simulations were excluded. The multivariate PSRF was between 1 and 1.01 in all of the cases, and the univariate diagnostics were between 1 and 1.01 in most cases. In some cases univariate diagnostics were somewhat higher – typically ν obtained the highest diagnostic value – but did not exceed 1.1 for any parameter.

Here r_{max} and s_{max} are maximum allowed lag and lead lengths, respectively.

We decided to estimate (32) using the reciprocal importance estimator of Gelfand and Dey (1994) because it is based on straightforward calculations and does not require the evaluation of the posterior density, which may be difficult in our case (see Geweke, 2005).⁴ It also seems to work well in practice as long as the truncation parameter M is not too large. For large M , the dimensionality of the parameter space may become too high, making the method inaccurate.

Given the marginal likelihoods $p(\mathbf{y}|M_{ij})$ of each of the models M_{ij} ($i = 0, \dots, r_{max}, j = 0, \dots, s_{max}$), model selection can be based on the posterior model probabilities. By assuming that our set of models is exhaustive, we have from Bayes' theorem that

$$p(M_{ij}|\mathbf{y}) = \frac{p(\mathbf{y}|M_{ij})p(M_{ij})}{\sum_{i=0}^r \sum_{j=0}^s p(\mathbf{y}|M_{ij})p(M_{ij})}, \quad (33)$$

where $p(M_{ij})$ is the prior model probability assigned to model M_{ij} . We assume that all the models are equally likely a priori because as long as r_{max} and s_{max} are reasonable, we have

⁴ A truncated multivariate normal density is used as the importance density. As suggested by Geweke (1999), the truncation is based on the $(1-p)$ th percentile of the chi-squared distribution with degrees of freedom equal to the number of elements in $(\boldsymbol{\varepsilon}^-, \boldsymbol{\varepsilon}^+, \boldsymbol{\theta})$ (see also Koop, 2003). In our applications $p = 0.05$ is chosen.

no reason to assume otherwise. That is, we set $p(M_{ij}) = 1/(r_{max}+1) \times (s_{max}+1)$ for all $i = 0, \dots, r_{max}$ and $j = 0, \dots, s_{max}$, and compute all the posterior model probabilities $p(M_{ij}|\mathbf{y})$. The model with the greatest posterior probability is selected.⁵ Lanne and Saikkonen (2008) suggest selecting the maximum lag and lead lengths by first finding a Gaussian $AR(r_{max}, 0)$ model with r_{max} sufficiently great to eliminate all serial correlation in the errors and then considering all $AR(r, s)$ models with $r + s = r_{max}$. Alternatively, information criteria could be employed.

To study the ability of Bayesian model selection in discriminating between causal and noncausal specifications, we conducted a small simulation experiment. Throughout, the results are based on 1,000 realizations of a series of 150 observations. To keep the number of simulations reasonable, we restrict our attention to a simple case where $r_{max} = s_{max} = 1$. Thus, the underlying data generating process is

$$(1 - \phi_1 L^{-1})(1 - \rho_1 L)y_t = \varepsilon_t, \quad (34)$$

where the error terms ε_t are assumed to have the standardized Student's t -distribution with 3 degrees of freedom and scale unity. The data are generated from (34) with various positive values of ρ_1 and ϕ_1 (given in Table 1 and 2). In order to reduce initialization effects, 100 observations at the beginning and end of each realization are discarded. For each realization we estimate four different models: a white noise model M_{00} ($r = 0, s =$

⁵ Notice that in case all the models are misspecified, that is, when none of them is the true model, the weight (posterior probability) of the best model in the Kullback-Leiber metric tends to 1 asymptotically (see Phillips, 1996; Fernández-Villaverde and Rubio-Ramírez, 2004).

0), a causal model M_{10} ($r = 1, s = 0$), a purely noncausal model M_{01} ($r = 0, s = 1$), and a noncausal model M_{11} ($r = 1, s = 1$).

The estimation is based on the posterior distribution (14), where the truncation parameter M is set at 30 and the joint prior density of Section 3 is used.⁶ Using demeaned data and the methods explained in the previous section, the posterior model probabilities $p(M_{11}|\mathbf{y})$ and $p(M_{01}|\mathbf{y})+p(M_{11}|\mathbf{y})$ are calculated, and averaged over 1,000 replications.⁷ The former quantity gives the mean posterior probability of the true noncausal model (except when $\phi_1 = 0$ or $\rho_1 = 0$), and the latter the mean posterior probability of a noncausal process, i.e., it can be interpreted as the overall probability of the presence of noncausality. Following Marriot and Newbold (2000), we also consider the decision rules $p(M_{11}|\mathbf{y}) > 0.5$ and $p(M_{01}|\mathbf{y})+p(M_{11}|\mathbf{y}) > 0.5$, indicating that a model is selected if its posterior model probability exceeds 50%. The means of the posterior model probabilities are presented in the upper panels and the proportions of times when $p(M_{11}|\mathbf{y}) > 0.5$ or $p(M_{01}|\mathbf{y})+p(M_{11}|\mathbf{y}) > 0.5$ are reported in the lower panels of Tables 1 and 2, respectively.

⁶ According to our experiments (not reported) $M = 30$ is large enough to guarantee a sufficiently accurate approximation to the joint density of observed data. In particular, as long as the zeros of the polynomials $\phi(z)$ and $\rho(z)$ are not close to the unit circle, α_j and β_j seem to be close enough to zero for $j > 30$, in the sense that the posterior model probabilities or the forecasting accuracy do not considerably change by further increases of M .

⁷ The number of simulation rounds N was set at 12,000, and the first 2,000 simulations in each chain were excluded as a burn-in period.

As expected, the greater are the parameters ρ_1 and ϕ_1 , the greater is the probability of the true noncausal processes. Furthermore, the procedure seems to perform fairly well in discriminating between causality and noncausality, selecting a noncausal process in over 93% of the replicates whenever the true value of ϕ_1 is greater than or equal to 0.3 according to the 50% rule above. On the other hand, the risk of selecting a noncausal process when the true process is in fact causal ($\phi_1 = 0$) is at an acceptable level, being less than 15% in all cases.

Finally, we compare our Bayesian model selection procedure to that of Lanne and Saikkonen (2008). They strongly recommend using diagnostic checks to confirm the adequacy of the model suggested by the maximized likelihood criterion, but we ignore this step as it is difficult to incorporate into the simulation experiment. For simplicity, we consider the case where the autoregressive order is assumed to be known. In particular, we set $r + s$ at 2 and calculate the marginal likelihoods and the maximum values of the approximate log likelihood function for the causal, purely noncausal and mixed models. We assume three parameter combinations, $(\rho_1, \phi_1) = \{(0.1, 0.7), (0.7, 0.1), (0.7, 0.7)\}$. In the first case, the data generating process (DGP) is close to purely noncausal, in the second case it is close to causal, and in the third case the roots of the lag polynomial are equal. Again, the results (not reported in detail) are based on 1,000 realizations of a series of 150 observations where the error terms ε_t are assumed to have the standardized Student's t -distribution with 3 degrees of freedom and scale unity. In general, the two procedures yield rather similar results although the Bayesian criterion slightly outperforms the classical procedure. For instance, the Bayesian criterion selects the true model in 67.4%, 65.6% and 99.8% of the replicates when (ρ_1, ϕ_1) are $(0.7, 0.1)$, $(0.1,$

0.7), and (0.7, 0.7), respectively. The corresponding classical figures are 60.7%, 61.6% and 96.5%. Furthermore, when the data are generated from a purely causal AR(2,0) process with $\rho_1 = 0.6$ and $\rho_2 = 0.2$, the classical procedure selects the noncausal model in 14.5% and the Bayesian procedure in 7.0% of the replicates. Thus, the risk of selecting a noncausal process when the true process is in fact causal tends to be higher in the classical procedure.

6. Empirical application

Today, one of the most interesting macroeconomic phenomena is the U.S. inflation. Questions concerning the true nature of inflation dynamics and the extent to which it can be forecast still remain without solid answers (see, e.g., Rudd and Whelan, 2006; Stock and Watson, 2007, 2008). Therefore, the methods introduced above are applied to the U.S. consumer price inflation. For the most part we restrict our attention to the question of whether the observed correlation between current and lagged inflation should be interpreted as evidence of price-setters' dependence on past inflation in addition to or instead of future expectations, but we also consider forecasting inflation.

6.1. Posterior results

The specific inflation series we study is the annualized quarterly inflation computed from the seasonally adjusted U.S. consumer price index (for all urban consumers) compiled by

the Bureau of Labor Statistics.⁸ The sample period covers 148 observations from 1970:1 to 2006:4. There is a substantial literature examining the behavior of this series (from different sample periods). The series is found to be highly persistent, measured in terms of serial correlation, which has typically been interpreted as evidence in favor of price-setters' depending on past inflation in forming expectations. However, since causal and purely noncausal processes can have the same autocorrelation function, high autocorrelation may arise even if the process is purely noncausal with no dependence on past values. Therefore, we use posterior model probabilities to find out about the true dynamics of the inflation process. The presence of noncausality in the same series has previously been studied by Lanne and Saikkonen (2008), who selected a purely noncausal AR(0,3) model.

As our data are quarterly, we set the maximum lag and lead lengths at four. Prior to estimation, the inflation series is demeaned. The posterior probabilities of the different AR(r,s) models are shown in the upper panel of Table 3.⁹

⁸ If the seasonal filter applied to the inflation data is noncausal, one would expect to find support for noncausality, even when the data are not generated from a noncausal process. Therefore, as a check, we also calculated marginal likelihoods for the different causal and noncausal models using the corresponding non-seasonally adjusted inflation series. The conclusions remained intact irrespective of the inflation series used (in both cases the data lend strong support to noncausality). The detailed results (not reported) are available upon request.

⁹ The posterior estimates of ϵ^- , ϵ^+ and θ are based on 100,000 draws. The first 10,000 draws are discarded as a burn-in period and the convergence of these chains is checked using the PSRFs.

Most interestingly, there is strong support in the data for a noncausal process; the probability of s being zero is less than 2%. In other words, the posterior probability of noncausality is over 98%. The model with the greatest posterior probability is the purely noncausal AR(0,3) model. Thus, the posterior probabilities indicate the same model as the model selection procedure of Lanne and Saikkonen (2008). Of course, the posterior probability of this particular model being the true model is low, only 18%, and the probability of the purely noncausal process ($r = 0$) is only 29.9%. Thus, as the prior probability of the purely noncausal process is 20% we have not learnt much about $r = 0$. These results suggest that U.S. inflation likely depends on its past values also, which could follow from some agents' basing their price-setting decisions on past inflation. Such a dependence on past inflation is in line with the substantial empirical literature concerning the rule-of-thumb behavior of producers and consumers (see, e.g., Campbell and Mankiw, 1990; Galí and Gertler, 1999; Galí *et al.* 2005; Smets and Wouters, 2007).

The lower panel of Table 3 reports the posterior summary statistics of the maximum a posteriori model, and Figure 1 shows the corresponding marginal posterior densities. The data appear to be particularly informative for these parameters. That is, the variances of the marginal posterior distributions are found to be small (with one exception, the degrees of freedom parameter ν)¹⁰, and, as a matter of fact, systematically smaller than the prior variances (see Figure 1). The marginal priors seemed to have only negligible influence on the marginal posteriors. The fact that our posterior results are very similar to the maximum likelihood estimates of Lanne and Saikkonen (2008) also attests to this. However, the posterior median of the degrees of freedom parameter ν (4.3) differs from

¹⁰ Notice, however, that the results are not sensitive to the choice of the prior mean of ν .

its ML estimate (3.7) because its posterior density is skewed to the right. From the graph of the marginal posterior density of v in Figure 1 the posterior mode is seen to be fairly close to its ML estimate.

6.2. Forecasting exercise

As pointed out above, forecasts are easily obtained as a by-product of Bayesian estimation of noncausal AR models. Out-of-sample forecasting is one of the major uses of univariate time series models, and, therefore, it is of interest to compare the forecasting performance of causal and noncausal AR models. Forecast horizons from one to four quarters are considered ($h = 1, 2, 3, 4$).

We employ the following recursive forecasting procedure for the Bayesian AR models: $(r_{max}+1) \times (s_{max}+1)$ different models, denoted M_{ij} ($i = 0, \dots, r_{max}$ and $j = 0, \dots, s_{max}$) are estimated on demeaned data from the in-sample period ($t = 1, \dots, T$), their marginal likelihoods are calculated, and the predictive densities over $T+1, \dots, T+h$ are computed using four alternative model selection strategies described below. Moving forward, the data for the period $t = 1, \dots, T+1$ are demeaned, all models are re-estimated, their marginal likelihoods are calculated and the predictive densities over $T+2, \dots, T+h+1$ are computed. This is continued until the end of the series. The period over which the dynamic forecast distributions are computed in this manner is 1984:1 through 2006:3.¹¹

¹¹ In the recursive forecast exercise, a total of 94 chains for each combination of s and r are simulated ($r_{max} = s_{max} = 4$, thus, altogether 2350 chains). The posterior estimates of ϵ^- , ϵ^+ and θ are based on 100,000 draws, and the first 10,000 draws are discarded as a burn-in period.

Several alternative strategies are applied in model selection at each step. In the first three approaches, posterior model probabilities are used to select the optimal lead and (or) lag lengths, and forecasts are generated from the selected model. The models are referred to as $AR(r_m, 0)$, $AR(0, s_m)$, when only purely causal and noncausal AR models are considered, respectively, and $AR(r_m, s_m)$ when all AR models are considered. In the fourth strategy, the effect of model uncertainty is controlled by Bayesian model averaging (BMA). In particular, the posterior predictive density of the future observation y_{T+h} is an average of the posterior predictive densities of y_{T+h} of all the models being considered, weighted by their posterior model probabilities (see, e.g., Hoeting *et al.*, 1999). In other words, the predictive density is obtained as

$$p(y_{T+h}|\mathbf{y}) = \sum_{j=1}^{(r_{\max}+1) \times (s_{\max}+1)} p(y_{T+h}|\mathbf{y}, M_{ij}) p(M_{ij}|\mathbf{y}). \quad (35)$$

This approach obviously takes into account the uncertainty in model selection by marginalizing out the unknown models (in our case the quantities r and s). Finally, third-order $AR(3,0)$ and $AR(0,3)$ models are used to compute forecasts at each step. In all approaches the posterior mean forecasts are used as a point forecasts. The results based on the posterior medians (not reported) turned out to be very close to those based on posterior means.

The predictive performance is measured by the root mean squared forecast error (RMSE). For comparison, we also consider the classical causal AR model, estimated by OLS, with the order selected by the Akaike Information Criterion (AIC) at each step, and refer to it

as the $AR(r_{aic})$ model. Following the standard practice in the literature on inflation forecasting (for a survey of recent literature, see Stock and Watson, 2008), we consider point forecasts of average inflation, given by $(400/h) \times \ln(P_{T+h}/P_T)$, where P_T is the value of the consumer price index at time T . However, since the model ranking based on the mean squared forecast criteria (MSE) is not invariant to nonsingular, scale-preserving linear transformations (see, e.g., Clements and Hendry, 1993), the results are checked by also forecasting a particular quarter in the future, that is, $400 \times \ln(P_{T+h}/P_{T+h-1})$.

Table 5 shows the RMSEs of the forecasting models relative to the benchmark naïve model (see, e.g., Atkeson and Ohanian, 2001).¹² The results of the causal and noncausal models with the maximum posterior probabilities, that is, the $AR(3,0)$ and $AR(0,3)$ models, and the results of the $AR(0,3)$ and $AR(0, s_m)$ models with $\Phi_0 = I$ (see the discussion of priors in Section 3), are also reported. In addition to the entire out-of-sample period, the forecasts are compared for the 1984:1 - 1994:2 and 1994:3 - 2006:4 subsample periods of equal length. This allows us to control for the possible structural break in the mid 1980s. The $AR(0,3)$ forecasts perform relatively well in all out-of-sample periods and at all forecasting horizons. However, the differences in favor of the purely noncausal models seem to increase with the forecast horizon, and at horizons of three and four quarters the $AR(0,3)$ model is always the best. In the entire out-of-sample period, the $AR(0,3)$ model produces the most accurate average inflation forecasts at every horizon. However, as Stock and Watson (2007), among others, have pointed out, there is a lot of

¹² When the average inflation is forecasted the naïve forecast is given by $(400/h) \times \ln(P_T/P_{T-h})$, whereas when a particular future quarter is forecasted it is given by $400 \times \ln(P_T/P_{T-1})$.

sampling variation in RMSEs of postwar U.S. inflation forecasts reported in the previous literature. In view of this it is not surprising that the differences in the RMSEs in Table 5 are relatively small (although clearly distinct).

Finally, the purely noncausal models with $\Phi_0 = I$ turn out to forecast better than those with shrinkage priors. However, especially in the case of AR(0,3), the difference between the results is negligible. Interestingly, model averaging performs poorly in all periods and horizons yielding less accurate forecasts than the competitive models.

7. Conclusion

In this paper, we have introduced Bayesian techniques to analyze the dynamics of economic time series in which expectations play an important role. In particular, we have studied the noncausal AR models, which allow for dependence on future as well as past values of the variable in question, from the perspective of model selection. In addition, we have shown how to generate forecasts from noncausal AR models in a straightforward manner.

We examined the finite-sample properties of our model selection procedure by means of simulation experiments. In general, our results indicate that the Bayesian posterior model probability criterion is able to discriminate between causal and noncausal specifications fairly well. In particular, the posterior probability in favor of a true noncausal process turned out to be high even with low degree of noncausality. Furthermore, when the true data generating process is causal, the Bayesian criterion selects the noncausal model

markedly less frequently than the classical procedure of Breidt *et al.* (1991) and Lanne and Saikkonen (2008). Thus, it seems that the probability of falsely selecting a noncausal process is lower with the Bayesian criterion.

The methods introduced in this paper were applied to U.S. consumer price inflation. The results show that the observed persistence in inflation is caused by the predictability of inflation. However, we also found evidence of the agents' price-setting behavior being dependent on past inflation, which contradicts typical New Keynesian models with forward-looking dynamics, but accords with much of the recent empirical literature. Finally, our forecasting results indicate that the forecasts based on purely noncausal AR models are, in general, more accurate than those based on causal AR models or noncausal AR models containing lagged terms, especially at longer forecast horizons.

References

Andrews B, Davis R A, Breidt F J. 2006. Maximum likelihood estimation for all-pass time series models. *Journal of Multivariate Analysis* **97**: 1638–1659.

DOI:10.1016/j.jmva.2006.01.005

Atkeson, A, Ohanian, L E. 2001. Are Phillips Curves Useful for Forecasting Inflation?.

Federal Reserve Bank of Minneapolis Quarterly Review **25**: 2–11.

Bartlett, M S. 1957. A comment on D.V. Lindley's statistical paradox. *Biometrika* **44**:

533–534. DOI:10.1093/biomet/44.3-4.533

Bauwens L, Lubrano M. 1998. Bayesian inference on GARCH models using Gibbs

sampler. *Econometrics Journal* **1**: 23–46. DOI: 10.1111/1368-423X.11003

Breidt J, Davis R A, Lii K S, Rosenblatt M. 1991. Maximum likelihood estimation for noncausal autoregressive processes. *Journal of Multivariate Analysis* **36**: 175–198.

DOI:10.1016/0047-259X(91)90056-8

Breidt J, Davis R A, Trindade A A. 2001. Least absolute deviation estimation for all-pass time series models. *Annals of Statistics* **29**: 919–946.

DOI:10.1214/aos/1013699987

Brockwell P A, Davis R A. 1991. *Time Series: Theory and Methods*, 2nd edition.

Springer-Verlag: New York.

Brooks S P, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**: 434–455.

DOI:10.1.1.70.2165

Campbell J Y, Mankiw N G. 1990. Permanent income, current income, and consumption. *Journal of Business and Economic Statistics* **8**: 265–79.

DOI:10.2307/1391964

Clements M P, Hendry D F. 1993. On the Limitations of Comparing Mean Squared Forecast Errors. *Journal of Forecasting* **12**: 617–637. DOI:10.1002/for.3980120802

Fernández-Villaverde J, Rubio-Ramírez J F. 2004. Comparing Dynamic Equilibrium Models to Data: A Bayesian Approach. *Journal of Econometrics* **123**: 153–187.

DOI:10.1016/j.jeconom.2003.10.031

Galí J, Gertler M. 1999. Inflation dynamics: A structural econometric analysis. *Journal of Monetary Economics* **44**: 195–222. DOI:10.1016/S0304-3932(99)00023-9

Galí J, Gertler M, López-Salido J D. 2005. Robustness of the estimates of the hybrid New Keynesian Phillips curve. *Journal of Monetary Economics* **52**: 1107–1118.

DOI:10.1016/j.jmoneco.2005.08.005

Gelfand A E, Dey D K. 1994. Bayesian model choice: Asymptotics and exact

calculations. *Journal of Royal Statistical Society Series B* **56**: 501–514.

Gelman A, Rubin D B. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* **7**: 457–472. DOI:10.1214/ss/1177011136

Geweke J. 1993. Bayesian treatment of the independent Student-*t* linear model. *Journal of Applied Econometrics* **8**: S19–S40. DOI:10.1002/jae.3950080504

Geweke J. 1999. Using simulation methods for Bayesian econometric models: Inference, development and communication. *Econometric Reviews* **18**: 1–126. DOI:10.1080/07474939908800428

Geweke J. 2005. *Contemporary Bayesian Econometrics and Statistics*. John Wiley & Sons: Hoboken, NJ.

Hansen L P, Sargent T J. 1991. Two difficulties in interpreting vector autoregressions. In *Rational Expectations Econometrics*, Hansen L P, Sargent T J (eds). Westview Press: Boulder, CO.

Hoeting J A, Madigan D, Raftery A E, Volinsky C T. 1999. Bayesian model averaging: A tutorial. *Statistical Science* **14**: 382–417. DOI:10.1214/ss/1009212519

Huang J, Pawitan Y. 2000. Quasi-likelihood estimation of noninvertible moving average processes. *Scandinavian Journal of Statistics* **27**: 689–702. DOI:10.1111/1467-

9469.00216

Kasa K, Walker T B, Whiteman C H. 2007. Asset prices in a time series model with perpetually disparately informed, competitive traders. Unpublished manuscript, Department of Economics, Simon Fraser University.

Koop G. 2003. *Bayesian Econometrics*. John Wiley & Sons: Chichester.

Lanne M, Saikkonen P. 2008. Modeling expectations with noncausal autoregressions. HECER Discussion Paper No. 212.

Litterman R B. 1986. Forecasting with Bayesian Vector Autoregressions: Five Years of Experience. *Journal of Business and Economic Statistics* **4**: 25–38.

Marriott J, Newbold P. 2000. The strength of evidence for unit autoregressive roots and structural breaks: A Bayesian perspective. *Journal of Econometrics* **98**: 1–25.

DOI:10.1016/S0304-4076(99)00077-9

O'Hagan A. 1995. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society Series B* **57**: 99–138.

Phillips, P C B. 1996. Econometric Model Determination. *Econometrica* **64**: 763–812.

Rudd J, Whelan K. 2006. Can rational expectations sticky-price models explain

inflation dynamics? *American Economic Review* **96**: 303–320.

DOI:10.1257/000282806776157560

Smets F, Wouters R. 2007. Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review* **97**: 586–606.

Stock J, Watson M. 2007. Why Has U.S. Inflation Become Harder to Forecast?.

Journal of Money, Credit and Banking **39**: 3–33. DOI:10.1111/j.1538-

4616.2007.00014.x

Stock J, Watson M. 2008. Phillips Curve Inflation Forecasts. NBER Working Paper No. W14322.

Tarami, B, Pourahmadi, M. 2000. Multi-variate t autoregressions: Innovations, prediction variances and exact likelihood equations. *Journal of Time Series Analysis* **24**: 739–754. DOI:10.1111/j.1467-9892.2003.00332.x

West, M. 1984. Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society Series B* **46**: 431–9.

Zellner, A. 1971. *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons: New York.

Table 1. Mean of $p(M_{11}|y)$ and proportion of times when $p(M_{11}|y) > 0.5$.Panel A. The mean of $p(M_{11}|y)$

ρ_1	ϕ_1					
	0.0	0.1	0.3	0.5	0.7	0.9
0.0	0.05	0.08	0.20	0.21	0.21	0.16
0.1	0.08	0.15	0.33	0.40	0.40	0.39
0.3	0.20	0.34	0.77	0.89	0.93	0.96
0.5	0.23	0.41	0.90	0.99	1.00	1.00
0.7	0.23	0.41	0.95	1.00	1.00	1.00
0.9	0.18	0.42	0.96	1.00	1.00	1.00

Panel B. Proportion of times when $p(M_{11}|y) > 0.5$

ρ_1	ϕ_1					
	0.0	0.1	0.3	0.5	0.7	0.9
0.0	0.00	0.02	0.06	0.07	0.08	0.05
0.1	0.02	0.05	0.22	0.33	0.33	0.32
0.3	0.07	0.23	0.80	0.93	0.96	0.98
0.5	0.10	0.35	0.93	1.00	1.00	1.00
0.7	0.10	0.34	0.97	1.00	1.00	1.00
0.9	0.07	0.36	0.98	1.00	1.00	1.00

The DGP is the AR(1,1) model where the error term follows the t -distribution with $\nu = 3$ and $\sigma = 1$. The results are based on 1,000 realizations of a series of 150 observations.

Table 2. Mean of $p(\text{Noncausality}|\mathbf{y})$ and proportion of times when $p(\text{Noncausality}|\mathbf{y}) > 0.5$.Panel A. The mean of the posterior probability $p(M_{01}|\mathbf{y})+p(M_{11}|\mathbf{y})$ of the noncausal process

ρ_1	ϕ_1					
	0.0	0.1	0.3	0.5	0.7	0.9
0.0	0.17	0.34	0.91	0.98	0.98	0.94
0.1	0.22	0.47	0.90	0.98	0.99	0.95
0.3	0.25	0.43	0.89	0.98	0.99	0.99
0.5	0.25	0.43	0.92	1.00	1.00	1.00
0.7	0.25	0.44	0.95	1.00	1.00	1.00
0.9	0.26	0.47	0.97	1.00	1.00	1.00

Panel B. Proportion of times when $p(M_{01}|\mathbf{y})+p(M_{11}|\mathbf{y}) > 0.5$

ρ_1	ϕ_1					
	0.0	0.1	0.3	0.5	0.7	0.9
0.0	0.03	0.25	0.93	0.99	0.99	0.96
0.1	0.07	0.42	0.94	0.99	0.99	0.98
0.3	0.13	0.35	0.93	0.99	1.00	1.00
0.5	0.13	0.38	0.95	1.00	1.00	1.00
0.7	0.13	0.38	0.97	1.00	1.00	1.00
0.9	0.15	0.42	0.99	1.00	1.00	1.00

The DGP is the AR(1,1) model where the error term follows the t -distribution with $\nu = 3$ and $\sigma = 1$. The results are based on 1,000 realizations of a series of 150 observations.

Table 3. Estimation results of the equation (1) for the U.S. inflation

Panel A. The posterior model probabilities $p(M_{ij}|\mathbf{y})$ for all combinations of r and s

r	s				
	0	1	2	3	4
0	0.000	0.000	0.001	0.181	0.087
1	0.000	0.000	0.001	0.121	0.064
2	0.000	0.013	0.005	0.057	0.060
3	0.013	0.030	0.165	0.023	0.050
4	0.005	0.024	0.021	0.050	0.031
$p(s = \bullet \mathbf{y})$	0.019	0.067	0.192	0.431	0.291

Panel B. Log marginal likelihoods $\ln p(\mathbf{y}|M_{ij})$ for all combinations of r and s

r	s				
	0	1	2	3	4
0	-388.28	-346.51	-335.47	-329.66	-330.40
1	-346.04	-339.76	-334.85	-330.06	-330.70
2	-340.27	-332.29	-333.31	-330.82	-330.76
3	-332.26	-331.45	-329.75	-331.72	-330.95
4	-333.18	-331.70	-331.83	-330.94	-331.42

Panel C. The point estimates of parameters of the maximum a posteriori model

Posterior/Par.	ϕ_1	ϕ_2	ϕ_3	ν	σ
Median	0.291	0.265	0.274	4.313	1.637
Std.	0.080	0.073	0.0726	2.286	0.181

All equations are estimated over the period 1970:1 to 2006:4

Table 4. The root mean squared forecast errors (RMSE) for the competitive models

	Relative RMSEs for $400 \times \ln(P_{T+h}/P_{T+h-1})$				Relative RMSEs for $(400/h) \times \ln(P_{T+h}/P_T)$			
Model	The out-of-sample forecasting period, 1984:1 - 2006:4							
-	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 1$	$h = 2$	$h = 3$	$h = 4$
AR(r_{aic})	0.807	0.799	0.928	0.858	0.807	0.957	1.054	1.152
AR($r_m, 0$)	0.778	0.784	0.907	0.830	0.778	0.928	1.018	1.108
AR(3, 0)	0.773	0.782	0.904	0.827	0.773	0.927	1.015	1.103
AR(0, 3)	0.766	0.791	0.896	0.815	0.766	0.929	0.994	1.073
AR(0, s_m)	0.781	0.800	0.913	0.837	0.781	0.949	1.022	1.112
AR(0, 3); $\Phi_0 = I$	0.765	0.788	0.892	0.810	0.765	0.924	0.987	1.063
AR(0, s_m); $\Phi_0 = I$	0.779	0.793	0.895	0.822	0.779	0.934	0.998	1.080
AR(r_m, s_m)	0.792	0.817	0.936	0.852	0.792	0.973	1.055	1.143
BMA	0.788	0.808	0.931	0.857	0.788	0.961	1.050	1.147
	The out-of-sample forecasting period, 1984:1 - 1994:2							
	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 1$	$h = 2$	$h = 3$	$h = 4$
AR(r_{aic})	0.915	0.840	0.969	0.897	0.915	0.974	1.037	1.150
AR($r_m, 0$)	0.853	0.814	0.948	0.864	0.853	0.952	0.999	1.108
AR(3, 0)	0.840	0.809	0.940	0.858	0.840	0.949	0.993	1.100
AR(0, 3)	0.832	0.806	0.945	0.857	0.832	0.943	0.995	1.099
AR(0, s_m)	0.862	0.814	0.963	0.885	0.862	0.966	1.020	1.135
AR(0, 3); $\Phi_0 = I$	0.834	0.801	0.937	0.849	0.834	0.942	0.989	1.089
AR(0, s_m); $\Phi_0 = I$	0.862	0.803	0.937	0.860	0.862	0.949	0.993	1.098
AR(r_m, s_m)	0.868	0.823	0.982	0.913	0.868	0.977	1.049	1.173
BMA	0.849	0.825	0.977	0.909	0.849	0.969	1.037	1.167
	The out-of-sample forecasting period, 1994:3 - 2006:4							
	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 1$	$h = 2$	$h = 3$	$h = 4$
AR(r_{aic})	0.750	0.771	0.900	0.826	0.750	0.914	1.075	1.155
AR($r_m, 0$)	0.740	0.763	0.878	0.803	0.740	0.908	1.042	1.107
AR(3, 0)	0.740	0.763	0.878	0.803	0.740	0.908	1.042	1.107
AR(0, 3)	0.733	0.780	0.861	0.780	0.733	0.916	0.992	1.038
AR(0, s_m)	0.739	0.791	0.878	0.799	0.739	0.933	1.025	1.081
AR(0, 3); $\Phi_0 = I$	0.729	0.780	0.860	0.779	0.729	0.910	0.985	1.029
AR(0, s_m); $\Phi_0 = I$	0.736	0.787	0.866	0.791	0.736	0.921	1.004	1.056
AR(r_m, s_m)	0.754	0.812	0.903	0.802	0.754	0.969	1.064	1.101
BMA	0.757	0.796	0.897	0.815	0.757	0.955	1.066	1.122

Reported figures are the RMSEs of the forecasting models relative to the Naïve benchmark model. At each step, the AR(r, s) model is selected by the Akaike information criterion from among the causal AR models (AR(r_{aic})) or using posterior model probabilities from among the causal (AR($r_m, 0$)), noncausal (AR(r_m, s_m)) and purely noncausal (AR(0, s_m)) AR models. The maximum orders of the lag polynomials, r and s , are set at four. The figures in the rows entitled AR(3, 0) and AR(0, 3) are based on the purely causal and noncausal third-order models, respectively, deemed the best in the entire sample. In the rightmost column entitled BMA, forecasts are computed by Bayesian model averaging of all the AR(r, s) models with $r, s = 0, 1, 2, 3, 4$. The bolded entries denote the lowest RMSE for period and horizon.

(Figure 1, J. Luoto)

Figure 1. Posterior densities of parameters of the maximum a posteriori model for the U.S. inflation (solid line) and their marginal prior densities (dots).