Bayesian Model Selection and Forecasting in Noncausal Autoregressive Models

Lanne, Markku and Luoma, Arto and Luoto, Jani

University of Helsinki, HECER, University of Tampere

September 2009

Helsinki Center of Economic Research

Discussion Papers

Bayesian Model Selection and Forecasting in Noncausal Autoregressive Models

Markku Lanne
University of Helsinki and HECER

Arto Luoma
University of Tampere

and

Jani Luoto
University of Helsinki and HECER

Discussion Paper No. 273
September 2009

ISSN 1795-0562

HECER – Helsinki Center of Economic Research, P.O. Box 17 (Arkadiankatu 7), FI-00014
University of Helsinki, FINLAND, Tel +358-9-191-28780, Fax +358-9-191-28781,
E-mail info-hecer@helsinki.fi, Internet www.hecer.fi

HECER
Discussion Paper No. 273

Bayesian Model Selection and Forecasting in Noncausal Autoregressive Models*

Abstract

In this paper, we propose a Bayesian estimation and prediction procedure for noncausal autoregressive (AR) models. Specifically, we derive the joint posterior density of the past and future errors and the parameters, which gives posterior predictive densities as a by-product. We show that the posterior model probability provides a convenient model selection criterion and yields information on the probabilities of the alternative causal and noncausal specifications. This is particularly useful in assessing economic theories that imply either causal or purely noncausal dynamics. As an empirical application, we consider U.S. inflation dynamics. A purely noncausal AR model gets the strongest support, but there is also substantial evidence in favor of other noncausal AR models allowing for dependence on past inflation. Thus, although U.S. inflation dynamics seem to be dominated by expectations, the backward-looking component is not completely missing. Finally, the noncausal specifications seem to yield inflation forecasts which are superior to those from alternative models especially at longer forecast horizons.

JEL Classification: C11, C32, C52, E31

Keywords: Noncausality, Autoregression, Bayesian model selection, Forecasting.

Markku Lanne

Department of Economics,
P.O. Box 17 (Arkadiankatu 7)
FI-00014 University of Helsinki
FINLAND

markku.lanne@helsinki.fi

Arto Luoma

Department of Mathematics and Statistics
FI-33014 University of Tampere
FINLAND

arto.luoma@uta.fi

Jani Luoto

Department of Economics,
P.O. Box 17 (Arkadiankatu 7)
FI-00014 University of Helsinki
FINLAND

jani.luoto@helsinki.fi

* Markku Lanne acknowledges financial support from the Okobank Group Research Foundation and the Academy of Finland.

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. 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.

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.[1] 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 determined on the basis of its autocorrelation structure because there are multiple causal and noncausal models of

[1] 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.

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 exact likelihood function where the past and future errors are considered unknown parameters. Thus the inference is not conditional on initial values. This is convenient in small samples, since in the conditional approach the number of starting values rapidly increases as a function of the orders of the autoregressive polynomials.

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 model selection criterion selects the noncausal model markedly less frequently than the procedure of Breidt et al. (1991) and Lanne and Saikkonen (2008). This indicates that the probability of falsely selecting the noncausal process is lower with our criterion.

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. The posterior medians of the purely noncausal AR model are also very close to those obtained by Lanne and Saikkonen (2008). Taken at face value, this finding indicates that the observed persistence in inflation is caused by the predictability of inflation instead of agents' relying on past inflation in forming expectations. However, even though the purely noncausal model turns out to be the likeliest by far, the probabilities of the other noncausal specifications, with dependence also on past inflation, are not negligible. This suggests that the persistence may be partly due to backward-looking behavior,

which 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.

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 compute forecasts. Moreover, in addition to employing a single model to produce forecasts, Bayesian model averaging is readily available. According to our out-of-sample forecasting exercise, the forecasts of inflation based on noncausal AR models turned out, in general, to be superior to those based on causal models, especially at longer forecast horizons. In most cases, the model suggested by our criterion produce the most accurate forecasts, only in the most recent subsample period did Bayesian model averaging produce the most accurate results.

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 results of the empirical application to the U.S. inflation. Finally, Section 7 concludes.

2. The model

Consider a stochastic process, y_t ($t = 0, \pm1, \pm2,\ldots$), generated by

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

where $\phi\left(L^{-1}\right) = 1 - \phi_1 L^{-1} - \cdots - \phi_s L^{-s}$, $\rho\left(L\right) = 1 - \rho_1 L - \cdots - \rho_r L^r$, ε_t is a sequence of i.i.d. random variables with zero mean and variance σ^2 and L is the lag operator. The autoregressive process defined in equation (1) is noncausal if $\phi_j \neq 0$ for some $j \in \{1,\ldots,s\}$ and it is referred to as purely noncausal when $\rho_1 = \cdots = \rho_r = 0$. We refer to the noncausal AR model (1) as the AR(r,s) model

that has the conventional causal AR model as a special case when $s = 0$ (see Lanne and Saikkonen, 2008, and the references therein).

In equation (1), we assume that the roots of the equation $\rho(z) = 0$ lie outside the unit circle, and hence the process $\phi(L^{-1})\, y_t = u_t$ has the following moving average representation,

$$u_t = \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j}\,, \tag{2}$$

where $\alpha_0 = 1$ and the coefficients α_j decay to zero at a geometric rate as $j \to \infty$. Similarly, we assume that the roots of $\phi(z) = 0$ lie outside the unit circle, implying that the process $\rho(L)\, y_t = v_t$ has the following moving average representation, showing dependence on future errors,

$$v_t = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t+j}\,, \tag{3}$$

where $\beta_0 = 1$ and the coefficients β_j decay to zero at a geometric rate as $j \to \infty$. The process y_t itself has the two-sided moving average representation

$$y_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}\,, \tag{4}$$

where ψ_j is the coefficient of z^j in the Laurent series expansion of $\rho(z)^{-1}\phi(z^{-1})^{-1} \overset{def}{=} \psi(z)$. Thus, y_t is a stationary and ergodic process with finite second moments.

Lanne and Saikkonen (2008) studied the maximum likelihood (ML) estimation of the noncausal autoregressive model specified in equation (1). In particular, they derived an approximate likelihood function in which the first r and last s observations are treated as fixed initial values. In this paper, an alternative approach is suggested, where also these observations and, hence, past and future errors are explicitly modeled by treating them as unobserved missing observations or unknown parameters. This is likely to improve estimation and is particularly useful in model comparison and forecasting because all available information is explicitly included in the analysis. Furthermore, for

optimal prediction of a noncausal process, knowledge of future errors is required, and, therefore, our approach has the advantage of allowing for straightforward way of computing forecasts.

For the Bayesian analysis of the noncausal AR model, 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. A truncated joint density function of the data $\mathbf{y} = (y_1, y_2,\ldots,y_T)'$ and the past and future errors conditional on the vector of parameters $\theta = (\rho_1,\ldots,\rho_r, \phi_1,\ldots,\phi_s, \sigma, \nu')'$, can be expressed as

$$p\left(\boldsymbol{\varepsilon}^{-},\mathbf{y},\boldsymbol{\varepsilon}^{+}\middle|\theta\right) = p\left(\boldsymbol{\varepsilon}^{-}\middle|\theta\right)p\left(\boldsymbol{\varepsilon}^{+}\middle|\theta\right)p\left(\mathbf{y}\middle|\theta,\boldsymbol{\varepsilon}^{-},\boldsymbol{\varepsilon}^{+}\right), \tag{5}$$

where $\boldsymbol{\varepsilon}^{-} = (\varepsilon_{-M},\ldots,\varepsilon_{-1}, \varepsilon_0)$, $\boldsymbol{\varepsilon}^{+} = (\varepsilon_{T+1}, \varepsilon_{T+2},\ldots, \varepsilon_{T+M})$, ν is an additional parameter vector consisting of the parameters that determine the shape of the error distribution, and the truncation parameter M is a positive integer, chosen to be large enough for a sufficiently good approximation of the joint density of the data, the last term on the right-hand side of equation (5). The likelihood function can be obtained by integrating out the past and future errors from the joint density (5),

$$p\left(\mathbf{y}\middle|\theta\right) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} p\left(\boldsymbol{\varepsilon}^{-}\middle|\theta\right)p\left(\boldsymbol{\varepsilon}^{+}\middle|\theta\right)p\left(\mathbf{y}\middle|\theta,\boldsymbol{\varepsilon}^{-},\boldsymbol{\varepsilon}^{+}\right)d\,\boldsymbol{\varepsilon}^{-}d\,\boldsymbol{\varepsilon}^{+}\ . \tag{6}$$

In a general case, 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.

We start by deriving the joint density function of the errors $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}^{-}, u_1,\ldots,u_r, \varepsilon_{r+1},\ldots,\varepsilon_{T-s}, v_{T-s+1},\ldots,v_T, \boldsymbol{\varepsilon}^{+})$ and then obtain $p\left(\boldsymbol{\varepsilon}^{-},\mathbf{y},\boldsymbol{\varepsilon}^{+}\middle|\theta\right)$ by means of the change of variables theorem. The ultimate goal is to express the known joint density of errors $\varepsilon_1,\ldots,\varepsilon_T$ as a function of the given data $\mathbf{y}$. Because the errors $\boldsymbol{\varepsilon}^{-}$, $u_1,\ldots,u_r$, $v_{T-s+1},\ldots,v_T$, and $\boldsymbol{\varepsilon}^{+}$ are independent of $\varepsilon_{r+1},\ldots,\varepsilon_{T-s}$, as can be seen from equations (2) and (3), the joint density function of $\boldsymbol{\varepsilon}$ has the expression

$$p\left(\boldsymbol{\varepsilon}\middle|\theta\right) = p\left(\boldsymbol{\varepsilon}^{-},u_1,\ldots,u_r\middle|\,\theta\right)\left(\prod_{t=r+1}^{T-s} f_\sigma\left(\varepsilon_t\middle|\,\nu\right)\right)p\left(v_{T-s+1},\ldots,v_T,\boldsymbol{\varepsilon}^{+}\middle|\,\theta\right)$$

$$= \prod_{t=-M}^{0} f_\sigma(\varepsilon_t|\nu)\, p\left(u_1,\ldots,u_r\middle|\boldsymbol{\varepsilon}^-,\theta\right)\prod_{t=r+1}^{T-s} f_\sigma(\varepsilon_t|\nu)\, p\left(v_{T-s+1},\ldots,v_T\middle|\boldsymbol{\varepsilon}^+,\theta\right)\prod_{t=T+1}^{T+M} f_\sigma(\varepsilon_t|\nu), \quad (7)$$

where the error distribution is assumed to be non-Gaussian with density $f_\sigma(x) = \sigma^{-1}f(\sigma^{-1}x, \nu)$. 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 ν, non-Gaussian, and positive for all real numbers x and all permissible values of the ν.

Now, by change of variables from $(u_1,\ldots,u_r,\ \varepsilon_{r+1},\ldots,\varepsilon_{T-s},\ v_{T-s+1},\ldots,v_T)$ to $\mathbf{y}$, equation (5) can be expressed as

$$p\left(\boldsymbol{\varepsilon}^-,\mathbf{y},\boldsymbol{\varepsilon}^+|\theta\right) = \prod_{t=-M}^{0} f_\sigma(\varepsilon_t|\theta)\, p\left(\phi\left(L^{-1}\right)y_1,\ldots,\phi\left(L^{-1}\right)y_r\middle|\boldsymbol{\varepsilon}^-,\theta\right)\prod_{t=r+1}^{T-s} f_\sigma\left(\phi\left(L^{-1}\right)\rho(L)y_t|\theta\right)$$

$$\times\, p\left(\rho(L)y_{T-s+1},\ldots,\rho(L)y_T\middle|\boldsymbol{\varepsilon}^+,\theta\right)\prod_{t=T+1}^{T+M} f_\sigma(\varepsilon_t|\theta)\,|A|, \quad (8)$$

where $|A|$ is the Jacobian determinant of the linear transformation given by equations $u_1 = \phi\left(L^{-1}\right)y_1$, $u_2 = \phi\left(L^{-1}\right)y_2,\ldots,u_{T-s} = \phi\left(L^{-1}\right)y_{T-s}$, $v_{T-s+1} = \rho(L)\,y_{T-s+1},\ldots,v_T = \rho(L)\,y_T$ (see Lanne and Saikkonen, 2008, for a detailed derivation of the joint density of $\mathbf{y}$). As a final step, we write $p\left(\phi\left(L^{-1}\right)y_1,\ldots,\phi\left(L^{-1}\right)y_r\middle|\boldsymbol{\varepsilon}^-,\theta\right)$ and $p\left(\rho(L)y_{T-s+1},\ldots,\rho(L)y_T\middle|\boldsymbol{\varepsilon}^+,\theta\right)$ in terms of the errors ε_t. Recalling that $\rho(L)\,u_t = \phi\left(L^{-1}\right)v_t = \varepsilon_t$ and noticing that the Jacobian determinants of the transformations from $p\left(u_1,\ldots,u_r\middle|\boldsymbol{\varepsilon}^-,\theta\right)$ to $p\left(\varepsilon_1,\ldots,\varepsilon_r\middle|\boldsymbol{\varepsilon}^-,\theta\right)$, and from $p\left(v_{T-s+1},\ldots,v_T\middle|\boldsymbol{\varepsilon}^+,\theta\right)$ to $p\left(\varepsilon_{T-s+1},\ldots,\varepsilon_T\middle|\boldsymbol{\varepsilon}^+,\theta\right)$ are unity, equation (8) can be written as

$$p\left(\boldsymbol{\varepsilon}^-,\mathbf{y},\boldsymbol{\varepsilon}^+|\theta\right) = \prod_{t=-M}^{0} f_\sigma(\varepsilon_t|\theta)\prod_{t=1}^{r} f_\sigma\left(\rho(L)\,u_t\middle|\boldsymbol{\varepsilon}^-,\theta\right)\prod_{t=r+1}^{T-s} f_\sigma\left(\phi\left(L^{-1}\right)\rho(L)y_t|\theta\right)$$

$$\times \prod_{t=T-s+1}^{T} f_\sigma\left(\phi\left(L^{-1}\right)v_t\middle|\boldsymbol{\varepsilon}^+,\theta\right)\prod_{t=T+1}^{T+M} f_\sigma(\varepsilon_t|\theta)\,|A|, \quad (9)$$

where $u_1 = \phi(L^{-1})\, y_1, \ldots, u_r = \phi(L^{-1})\, y_r$ and $v_{T-s+1} = \rho(L)\, y_{T-s+1}, \ldots, v_T = \rho(L)\, y_T$, that is, $u_1,\ldots,u_r$ and $v_{T-s+1},\ldots,v_T$ are calculated from the data.

Evaluating the right-hand side of equation (9) requires knowledge of $u_{1-r},\ldots,u_0$ and $v_{T+1},\ldots,v_{T+s}$, which cannot be computed directly from the data, but they can be obtained, for example, by the following simple recursive calculations. From equation (1) we have

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

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

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

where $u_{-M-1},\ldots,u_{-M-r}$, 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 for the integral (6). Furthermore, as pointed out above, explicit incorporation of the past and future errors into the analysis allows for computing optimal forecasts from a noncausal AR model in a straightforward manner. Specifically, the posterior densities of v_{T+1}, $v_{T+2},\ldots,v_{T+M}$ may be simulated using equation (10) and the posterior distributions of θ and ε_{T+1}, $\varepsilon_{T+2},\ldots$, ε_{T+M}. Then the predictive densities of the future observables $y_{t+1},\ldots,y_{t+h}$ can be calculated using the recursive formula $y_{t+h} = \rho_1 y_{t+h-1} + \cdots + \rho_r y_{t+h-r} + v_{t+h}$. The means or medians of these predictive distributions can be used as point forecasts. Notice also that 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, $\rho_1,\ldots,\rho_r$, $\phi_1,\ldots,\phi_s$, σ and ν. We use proper priors for these parameters 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 O'Hagan, 1995). This is, of course, the case when autoregressive models, with the unknown orders of the autoregressive polynomial operators, are compared.

For the parameters $\rho_1,\ldots,\rho_r$ and $\phi_1,\ldots,\phi_s$ we adopt a multivariate Student prior, $\rho_1,\ldots,\rho_r,\phi_1,\ldots,\phi_s \sim t_{s+r}(\mu,1,P,v_0)$, with mean vector μ and covariance matrix $P^{-1}/(v_0-2)$ (see Bauwens, Lubrano, and Richard, 1999). This prior corresponds to the familiar conjugate normal-inverted gamma prior for linear regression models when the scale parameter of the inverted gamma distribution is set at unity. We assume that μ is a zero vector, $P = k\,I$, where I is an identity matrix, k is a scalar, and v_0 equals 3. We set k at unity, indicating an identity prior covariance matrix 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. Alternatively, we could use, for example, a multivariate normal prior with an identity covariance matrix.

In order to study how our multivariate Student prior affects the posterior, a small simulation experiment is carried out. Specifically, we compare the posterior modes based on the Student and normal priors and the ML estimator. The data are generated as follows. First, starting with $u_1 = \ldots = u_r = 0$, a series from the causal model $\rho(L)u_t = \varepsilon_t$ ($t = r+1,\ldots,T$) is generated. Then y_t is computed recursively from $\phi(L^{-1})y_t = u_t$ for $t = T-s,\ldots,1$ by setting $y_{T-s+1} = \ldots = y_T = 0$.[2] We follow Lanne and Saikkonen (2008) and assume from now on that the error term ε_t has Student's t-distribution with $v > 2$ degrees of freedom and variance σ^2. We set v at 3 and σ^2 at 1 and consider three different combinations of parameters values, $(\rho_1, \phi_1) = \{(0.1,0.7), (0.7,0.1), (0.7,0.7)\}$. 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. A gamma prior with mean and

[2] In order to reduce initialization effects, 100 observations at the beginning and end of each realization are discarded.

variance set at unity is adopted for σ and an exponential prior for $\nu - 2$, where the mean of ν is set at 3 (these priors are explained in detail below). The results are based on a series of 150 observations (the time series used in our empirical application in Section 6 consist of 148 observations) and 1,000 replications. Table 1 presents the means and standard deviations of the posterior mode estimates of ρ_1 and ϕ_1 based on both the Student and normal priors as well as the means and standard deviations of the ML estimators obtained using the approximate likelihood function of Lanne and Saikkonen (2008). As one would expect, the differences between the ML and posterior results are minor. Thus, in this sense, both of these priors have only negligible influence on the posterior inference. The results based on several different DGPs (not reported) are also similar, but it seems that the posterior density is affected more strongly by the multivariate normal prior than by the Student prior (with $\nu_0 = 3$) when the number of lags (and leads) increases. Therefore, we recommend using the Student prior.

In order to seek for a suitable prior for σ, we estimated several posterior distributions of θ using different priors for σ, the full likelihood function (9) and the priors (for other parameters) and the artificial data used above. Our estimation results (not reported) indicate that the posterior distribution of σ has a long right tail. Therefore, we recommend using a tight prior on σ to facilitate numerical maximization. A gamma prior with mean and variance set at unity seems to work well in our experiments and it is the prior adopted in our simulation study in section 5 and empirical application in Section 6.

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. For computational reasons we give an exponential prior for $\nu - 2$ (instead of ν) and, in our simulation study, set the prior mean of ν at 3, which implies a tight prior variance, unity. In the empirical application we give more weight to the data and set the prior mean of ν at 7.

4. Posterior analysis

With the likelihood function derived in Section 2 and the prior distributions specified in Section 3, we are able to compute the posterior distribution of the parameters θ and the past and future unobserved errors. Let $p(\theta)$ and $p(\mathbf{y})$ denote the joint prior density of the parameters obtained as the

product of the marginal prior densities of the previous section and the joint density of the data, respectively. The joint posterior density of $\boldsymbol{\varepsilon}^-$, $\boldsymbol{\varepsilon}^+$ and θ can then be expressed as

$$p\left(\theta,\boldsymbol{\varepsilon}^-,\boldsymbol{\varepsilon}^+|\mathbf{y}\right)=\frac{p(\theta)p\left(\boldsymbol{\varepsilon}^-,\mathbf{y},\boldsymbol{\varepsilon}^+|\theta\right)}{p(\mathbf{y})}$$
$$=\frac{p(\theta)p\left(\boldsymbol{\varepsilon}^-,\mathbf{y},\boldsymbol{\varepsilon}^+|\theta\right)}{\int\cdots\int p(\theta)p\left(\boldsymbol{\varepsilon}^-,\mathbf{y},\boldsymbol{\varepsilon}^+|\theta\right)d\,\boldsymbol{\varepsilon}^-d\,\boldsymbol{\varepsilon}^+d\,\theta}\ . \tag{13}$$

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. Because none of the full conditionals of the density (13) are in the form of any standard probability density function (p.d.f.), we apply the Metropolis-Hastings algorithm.

We first apply the logarithmic transformations for the parameters σ and ν to obtain approximate normality of their marginal posteriors, which makes the posterior simulations markedly more efficient. Let $\eta = (\rho_1,\ldots,\rho_r,\phi_1,\ldots,\phi_s,\ \ln\sigma,\ \ln(\nu-2))'$ denote the vector of the transformed parameters. As starting values we use a zero vector for $\boldsymbol{\varepsilon}^-$ and $\boldsymbol{\varepsilon}^+$ and the posterior mode for η.[3]

In the ith iteration ($i = 1,\ldots,N$) we draw a candidate η^* from the normal proposal density and accept it with probability

$$\alpha=\min\left\{1,\frac{p\left(\eta^*\right)p\left(\boldsymbol{\varepsilon}^-_{(i-1)}|\theta^*\right)p\left(\boldsymbol{\varepsilon}^+_{(i-1)}|\theta^*\right)p\left(\mathbf{y}|\theta^*,\boldsymbol{\varepsilon}^-_{(i-1)},\boldsymbol{\varepsilon}^+_{(i-1)}\right)}{p\left(\eta_{(i-1)}\right)p\left(\boldsymbol{\varepsilon}^-_{(i-1)}|\theta_{(i-1)}\right)p\left(\boldsymbol{\varepsilon}^+_{(i-1)}|\theta_{(i-1)}\right)p\left(\mathbf{y}|\theta_{(i-1)},\boldsymbol{\varepsilon}^-_{(i-1)},\boldsymbol{\varepsilon}^+_{(i-1)}\right)}\right\}. \tag{14}$$

[3] To get a convenient proposal density for θ, we minimize the negative of the logarithm of the posterior density numerically (using the approximate likelihood function of Lanne and Saikkonen, 2008) to obtain the posterior mode of the transformed parameter vector η and evaluate the Hessian matrix at the minimum. We then compute the inverse of the Hessian to approximate the posterior covariance matrix of η and scale it by the factor $2.4^2/(s+r+2)$ to obtain the optimal covariance matrix Σ for the multivariate normal proposal distribution $f_N(\eta|\eta_{(i-1)},\Sigma)$ ($i = 1,\ldots,N$). Notice that the greater is the number of iterations N, the more precise are the posterior estimates. In some cases the covariance matrix estimate based on the local behavior of the posterior at its highest peak gives too optimistic a view of precision and thus fails to yield an efficient covariance matrix for the normal proposal distribution. In these cases, starting with the inverse of the Hessian in the proposal distribution we first simulate a certain number of posterior draws, use them to estimate $Cov(\theta|\mathbf{y},\boldsymbol{\varepsilon}^-,\boldsymbol{\varepsilon}^+)$, and then set $\Sigma = 2.4^2 Cov(\theta|\mathbf{y},\boldsymbol{\varepsilon}^-,\boldsymbol{\varepsilon}^+)/(r+s+2)$ (see e.g. Gelman et al., 2004). This is repeated until a reasonable value of Σ is found.

If η^* is not accepted, we set $\eta_{(i)}$ at its current value. The one-to-one mapping between η and θ is used.[4]

We continue by drawing a candidate $\boldsymbol{\varepsilon}^{-*}$ using $p(\boldsymbol{\varepsilon}^-|\theta_{(i)}) = f_\sigma(\varepsilon_{-M}|\theta_{(i)})\cdots f_\sigma(\varepsilon_0|\theta_{(i)})$ as the candidate generating density, and hence by the definition of the acceptance probability of the Metropolis-Hastings algorithm, the acceptance probability of this step simplifies to

$$\alpha = \min\left\{1, \prod_{t=1}^{r} f_\sigma\left(\rho(L)\,u_t\middle|\boldsymbol{\varepsilon}^{-*}, \theta_{(i)}\right) \Big/ \prod_{t=1}^{r} f_\sigma\left(\rho(L)\,u_t\middle|\boldsymbol{\varepsilon}^{-}_{(i-1)}, \theta_{(i)}\right)\right\} . \tag{15}$$

We set $\boldsymbol{\varepsilon}^{-}_{(i)} = \boldsymbol{\varepsilon}^{-*}$ with probability α and $\boldsymbol{\varepsilon}^{-}_{(i)} = \boldsymbol{\varepsilon}^{-}_{(i-1)}$ with probability $1 - \alpha$. Similarly, we take a candidate draw $\boldsymbol{\varepsilon}^{+*}$ from density $p(\boldsymbol{\varepsilon}^+|\theta_{(i)}) = f_\sigma(\varepsilon_{T+1}|\theta_{(i)})\cdots f_\sigma(\varepsilon_{T+M}|\theta_{(i)})$ and calculate an acceptance probability

$$\alpha = \min\left\{1, \prod_{t=T-s+1}^{T} f_\sigma\left(\phi\left(L^{-1}\right)v_t\middle|\boldsymbol{\varepsilon}^{+*}, \theta_{(i)}\right) \Big/ \prod_{t=T-s+1}^{T} f_\sigma\left(\phi\left(L^{-1}\right)v_t\middle|\boldsymbol{\varepsilon}^{+}_{(i-1)}, \theta_{(i)}\right)\right\} . \tag{16}$$

Again we set $\boldsymbol{\varepsilon}^{+}_{(i)} = \boldsymbol{\varepsilon}^{+*}$ with probability α and $\boldsymbol{\varepsilon}^{+}_{(i)} = \boldsymbol{\varepsilon}^{+}_{(i-1)}$ with probability $1 - \alpha$.

When several alternative models are estimated by the method discussed above, typically model comparison is of interest, and it can conveniently be based on the marginal likelihoods of the alternative models. For each noncausal AR model specification, M_j (j = 1,...,J), the marginal likelihood is simply the denominator of the joint posterior density (13),

$$p\left(\mathbf{y}\middle|M_j\right) = \int\cdots\int p\left(\theta\middle|M_j\right)p\left(\boldsymbol{\varepsilon}^-, \mathbf{y}, \boldsymbol{\varepsilon}^+\middle|\theta, M_j\right)d\,\boldsymbol{\varepsilon}^-\,d\,\boldsymbol{\varepsilon}^+\,d\,\theta\ . \tag{17}$$

This can be estimated from the simulated posterior sample using, for example, the reciprocal importance estimator of Gelfand and Dey (1994), given by

[4] Notice that by change of variable the prior of $\bar{\nu} = \ln(\nu - 2)$ is given by $p(\bar{\nu}) = \lambda\exp\{-\lambda e^{\bar{\nu}} + \bar{\nu}\}$, where $-\infty < \bar{\nu} < \infty$ and λ is rate parameter. By similar reasoning the prior of $\bar{\sigma} = \ln\sigma$ is $p(\bar{\sigma}) = \Gamma(1)^{-1}\exp\{\bar{\sigma} - e^{\bar{\sigma}}\}$, where $-\infty < \bar{\sigma} < \infty$.

$$\hat{p}\left(\mathbf{y}\middle|M_j\right)=\left[\frac{1}{G}\sum_{g=1}^{G}\frac{f\left(\gamma_j^{(g)}\right)}{p\left(\mathbf{y}\middle|\gamma_j^{(g)},M_j\right)p\left(\gamma_j^{(g)}\middle|M_j\right)}\right]^{-1}, \tag{18}$$

where $\gamma_j = (\boldsymbol{\varepsilon}_j^-, \boldsymbol{\varepsilon}_j^+, \eta_j)'$ is the vector of all unobservable variables, $p(\mathbf{y}|\gamma_j,M_j)$ is the likelihood for model M_j defined on region Γ_j, $p(\gamma_j|M_j)$ is the corresponding prior density, $f(\gamma_j)$ is any p.d.f. with support contained in Γ_j and $\left\{\gamma_j^{(g)}\right\}_{g=1}^{G}$ is a sample of size G from the estimated joint posterior distribution. We decided to use this method because it is based on straightforward calculations and does not require the evaluation of the posterior density $p(\gamma_j|\mathbf{y},M_j)$, which may be difficult in our case. 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.

The asymptotic theory behind the method of Gelfand and Dey (1994) implies that $f(\gamma_j)/[p(\gamma_j|M_j)p(\mathbf{y}|\gamma_j,M_j)]$ must be bounded from above (see e.g. Koop, 2003). To verify that this quantity is finite for all possible values of γ_j, we follow Geweke (1999) and let $f(\gamma_j)$ be a truncated multivariate normal density

$$f\left(\gamma_j\right)=\frac{1}{p(2\pi)^{m/2}}\left|\hat{\Omega}_j\right|^{-1/2}\exp\left[-\frac{1}{2}\left(\gamma_j-\hat{\gamma}_j\right)'\hat{\Omega}_j^{-1}\left(\gamma_j-\hat{\gamma}_j\right)\right]1\left(\gamma_j\in\Gamma_j\right), \tag{19}$$

where $\hat{\gamma}_j$ and $\hat{\Omega}_j$ are estimates of the mean and covariance matrix of the posterior density, respectively. The indicator function 1(•) takes the value 1 when $\gamma_j \in \Gamma_j$, where

$$\Gamma_j=\left\{\gamma_j:\ \left(\gamma_j-\hat{\gamma}_j\right)'\hat{\Omega}_j^{-1}\left(\gamma_j-\hat{\gamma}_j\right)\le\chi^2_{1-p}\left(m_j\right)\right\}, \tag{20}$$

$\chi^2_{1-p}\left(m_j\right)$ is the (1–p)th percentile of the Chi-square distribution with m_j degrees of freedom, and m_j is the number of elements in γ_j. In our applications $p = 0.05$ seems to work fairly well.

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 by the data. The nonnestedness of the alternative AR(r,s) models complicates classical model selection, but poses no particular problems in Bayesian analysis.

Given the marginal likelihoods $p(\mathbf{y}|M_j)$ of each of the J models M_j ($j = 1,\ldots,J$) discussed in Section 4, 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

$$p(M_j|\mathbf{y}) = \frac{p(\mathbf{y}|M_j)p(M_j)}{\sum_{i=1}^{J} p(\mathbf{y}|M_i)p(M_i)}, \tag{21}$$

where $J = (r_{max}+1)\times(s_{max}+1)$, and r_{max} and s_{max} are maximum allowed lag and lead lengths, respectively, and $p(M_j)$ is the prior model probability assigned to model M_j. We assume that all the models are equally likely a priori because as long as r_{max} and s_{max} are reasonable, we have no reason to assume otherwise. That is, we set $p(M_j) = 1/J$ for all j, and seek the posterior model probabilities $p(M_j|\mathbf{y})$ of all the combinations of $r = 0,\ldots,r_{max}$ and $s = 0,\ldots,s_{max}$. The model with the greatest posterior probability is selected. 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 1000 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

$$(1-\phi_1 L^{-1})(1-\rho_1 L)y_t = \varepsilon_t, \tag{22}$$

where the error terms ε_t are assumed to have the standardized Student's t-distribution with 3 degrees of freedom and variance unity. The data are generated from (22) with various positive values of ρ_1 and ϕ_1 (given in Table 2 and 3). 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_1 ($r = 0$, $s = 0$), a causal model M_2 ($r = 1$, $s = 0$), a purely noncausal model M_3 ($r = 0$, $s = 1$), and a noncausal model M_4 ($r = 1$, $s = 1$).

The estimation is based on the posterior distribution (13), where the truncation parameter M is set at 20 and the joint prior density of Section 3 is used.[5] Using demeaned data and the methods explained in the previous section, the posterior model probabilities $p(M_4|\mathbf{y})$ and $p(M_3|\mathbf{y})+p(M_4|\mathbf{y})$ are calculated, and averaged over 1000 replications.[6] The former quantity gives the mean posterior probability of the true noncausal model (except when $\rho_1 = 0$), 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 $p(M_4|\mathbf{y}) > 0.5$ and $p(M_3|\mathbf{y})+p(M_4|\mathbf{y}) > 0.5$, 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 $p(M_4|\mathbf{y}) > 0.5$ or $p(M_3|\mathbf{y})+p(M_4|\mathbf{y}) > 0.5$ are reported in the lower panels of Tables 2 and 3, respectively.

As we would expect, the greater is the parameter ϕ_1 the greater is the probability of a noncausal processes. There is, however, one exception: When the true value of ϕ_1 is close to unity and ρ_1 is close to zero, the mean posterior probability of the noncausal process decreases sharply. Notice that this is not a unit root issue. Rather, the noncausal and causal models are indistinguishable when $\phi_1 = 1$ and $\rho_1 = 0$ (or when $\phi_1 = 0$ and $\rho_1 = 1$). Therefore, as ϕ_1 approaches 1 under $\rho_1 \approx 0$, the probability of incorrectly selecting the causal process increases sharply. Otherwise, the procedure seems to perform fairly well in discriminating between causality and noncausality, selecting a

[5] According to our simulation experiments (not reported) $M = 20$ is large enough to guarantee a sufficiently accurate approximation to the joint density of observed data.

[6] The number of simulation rounds N was set at 20000, and the first 2000 simulations in each chain were excluded as a burn-in period. The convergence of the chains was checked using the standard convergence diagnostic of Geweke (1992). To reduce the size of output files, every 9th draw is used in the calculation of marginal likelihoods, thus $G = 2000$.

noncausal process in over 85% of the replicates whenever the true value of ϕ_1 is greater than or equal to 0.3.

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 order of the autoregressive polynomial operators 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 the same three parameter combinations (ρ_1, ϕ_1) = {(0.1,0.7), (0.7,0.1), (0.7,0.7)} as in section 3. Again, the results (not reported in detail) are based on 1000 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 variance unity. In general, the two procedures yield rather similar results although there are some differences. For instance, compared to the classical procedure, the Bayesian criterion performs very well when the process is clearly noncausal, selecting the true model in 76.7% and 98.4% of the replicates when (ρ_1, ϕ_1) = (0.1, 0.7) and (ρ_1, ϕ_1) = (0.7, 0.7), respectively. The corresponding classical figures are 67.9% and 96.5%. In contrast, when (ρ_1, ϕ_1) = (0.7, 0.1), the classical procedure selects the true model in 70.1% of the replicates, while the Bayesian procedure only reaches 59.1%. However, in this case, the true model is fairly close to the first-order causal autoregressive model. Therefore, this result may reflect the tendency of the classical procedure to select a noncausal process too frequently. Actually, when the data are generated from a purely causal AR(2,0) process where $\rho_1 = 0.6$ and $\rho_2 = 0.2$, the classical procedure selects the noncausal model in 12.4% and the Bayesian procedure in 8.1% of the replicates.

6. Empirical application

Today, one of the most interesting macroeconomic phenomena is the U.S. inflation. Questions like whether it is forward- or backward-looking or why it is so difficult to forecast it well have remained without solid answers (see e.g. Rudd and Whelan, 2006, and Stock and Watson, 2007). Therefore, the methods introduced in this paper 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 backward-looking inflation, 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). The series is published by the Bureau of Labor Statistics. 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 been interpreted as evidence in favor of backward-looking behavior of price setters. However, since causal and purely noncausal processes can have the same autocorrelation functions, the backward-looking or forward-looking behavior cannot be discriminated by this measure. We will therefore use posterior model probabilities in studying the role of forward-looking behavior in 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 4.[7] Most interestingly, there is strong support in the data for a noncausal process; the probability of s being zero is only 3.2%. In other words, the posterior probability of noncausality is 96.8%. 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). This purely noncausal model suggests that the inflation process is driven by expectations of future errors, whose predictability makes the series persistent. However, the posterior probability of this particular model being the true model is relatively low, approximately 28%, and the probability of purely noncausal process (r = 0) is only 36%. This suggests that U.S. inflation might to some extent depend on its past values, which could follow from some agents using a backward-looking rule to set prices. Such a dependence on past inflation would be in line with the substantial empirical literature concerning the rule-of-thumb behavior of

[7] The posterior estimates of $\boldsymbol{\varepsilon}^-$, $\boldsymbol{\varepsilon}^+$ and θ are based on 50,000 draws. The first 10,000 draws are discarded as a burn-in period and the convergence of these chains is checked using the convergence diagnostic of Geweke (1992).

producers and consumers (see e.g. Campbell and Mankiw, 1990, Galí and Gertler, 1999, Galí et al. 2005, and Smets and Wouters, 2007).

The lower panel of Table 4 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 in all these parameters. That is, the variances of the marginal posterior distributions are found to be systematically smaller than the prior variances. The marginal priors seem to have only negligible influence on the marginal posteriors. The fact that our results are very similar to the maximum likelihood (ML) estimates of Lanne and Saikkonen (2008) also attests to this. However, the posterior median of the degrees of freedom parameter ν (4.5) differs from its ML estimate (3.7) because its posterior density is skewed to the right. From the graph of the marginal posterior density of ν 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 both noncausal AR models and a number of other models suggested for inflation series in the previous literature. 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: J different models (j = 1,…,J) are estimated on demeaned data from the in-sample period (t = 1,…,T), their marginal likelihoods are calculated, and, using equation (10), the predictive densities over $y_{T+1},\ldots,y_{T+h}$ are computed using two alternative model selection strategies described below. Moving forward, the data for period t = 1,…,T+1 are demeaned, all models are re-estimated, their marginal likelihoods are calculated and the predictive densities over $y_{T+2},\ldots,y_{T+h+1}$ are computed. This is continued until the end of our time series. The period over which the dynamic forecast distributions are computed in this manner is 1984:1 through 2006:3.[8] This period is particularly interesting

[8] In the recursive forecast exercise, a total of 92 chains for each combination of s and r are simulated (r_{max} = s_{max} = 4, thus, altogether 2300 chains). The posterior estimates of $\boldsymbol{\varepsilon}^-$, $\boldsymbol{\varepsilon}^+$ and θ are based on 50,000 draws, and the first 10,000 draws are discarded as a burn-in period. To reduce the size of output files, every 4th draw is used in the calculation of

because, according to Atkeson and Ohanian (2001) and Stock and Watson (2007), for the U.S. the forecasts based on standard multivariate forecasting methods have been found inferior to the naïve forecast based on the univariate random walk since 1984.

Two alternative strategies are applied in model selection at each step. In the first approach, posterior model probabilities are used to select the optimal lead and lag lengths, and forecasts are generated from the selected model, referred to as AR(r_m, s_m). In the alternative 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 the J 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

$$p\left(y_{T+h}|\mathbf{y}\right)=\sum_{j=1}^{J} p\left(y_{T+h}\left|\mathbf{y},M_j\right.\right)p\left(M_j|\mathbf{y}\right). \tag{23}$$

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). In both approaches the posterior median forecasts are used as a point forecasts. The results based on the posterior means (not reported) turned out to be very close to those based on posterior medians.

The predictive performance of noncausal and causal Bayesian models is compared to that of the naïve model, the classical first-order integrated moving average (IMA(1,1)) model, suggested by Nelson and Schwert (1977) for inflation, and the conventional classical causal autoregressive model where the optimal lag length is determined by the Akaike Information Criterion (AIC). The naïve forecast we consider is given by the average annualized quarterly inflation over the previous h quarters. The IMA(1,1) model, in turn, is given by

$$\Delta y_t=\left(1-\omega L\right)\eta_t, \tag{24}$$

where ω is a parameter to be estimated and η_t is a normally distributed error term (see Stock and Watson, 2007, and the references therein). As Nelson and Schwert (1977) and Fama and Gibbons (1984), among many others, point out, the sample autocorrelations of U.S. inflation suggest a first-

---

the marginal likelihoods and every 20th draw is used in the calculation of the predictive distributions. The convergence of these chains is checked using the convergence diagnostic of Geweke (1992).

order moving average process as a model for Δy_t, as given in equation (24). On the other hand, because the first autocorrelation coefficient of y_t estimated from our data is roughly 0.6 and the higher-order coefficients decline slowly, a causal AR(r,0) model could be a suitable time series model for the level of inflation. We use the Akaike Information Criterion (AIC) to select the order of this classical causal AR(r,0) model at each step and refer to it as the AR(r_{aic},0) model.

The commonly used measure of forecasting accuracy, the root mean squared forecast error (RMSE) is applied for the h-step-ahead forecast errors $e_T(h) = y_{T,T+h} - y_{T,T+h|T}$, where $y_{T,T+h|T}$ is a point forecast of $y_{T,T+h}$ (such as posterior median), and $y_{T,T+h}$ is the average annualized quarterly inflation over the h quarters. Table 5 shows the RMSEs of the forecasting models relative to the benchmark AR(r_{aic},0) model. 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 out-of-sample forecasts indicate the superiority of noncausal AR models over the alternatives. The AR(r_m, s_m) forecast performs very well in all out-of-sample periods and at all forecasting horizons. However, the differences in favor of the noncausal model seem to increase with the forecast horizon. Interestingly, model averaging performs noticeably better only in the most recent period, having the lowest RMSEs at all forecasting horizons. Otherwise, it yields similar or slightly less accurate forecasts than the AR(r_m,s_m) model. Finally, improvement of the one-year-ahead naïve forecasts over the IMA(1,1) and AR(r_{aic},0) forecasts is smaller in our paper than in Stock and Watson (2007), which suggests that our benchmark models are well specified.

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. However, although Bayesian model selection works well, it has difficulties in discriminating between causal and noncausal specifications when the true model is a first-order causal or purely noncausal process with a near unit root, in which case the noncausal and causal models are almost indistinguishable.

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 instead of economic agents' relying on past inflation in forming expectations. However, we also found some evidence of the backward-looking behavior of agents, which contradicts typical New Keynesian models with forward-looking dynamics, but accords with much of recent empirical literature. Finally, our forecasting results indicate that the forecasts based on noncausal AR models are, in general, superior to those based on causal models, especially at longer forecast horizons.

References

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

Atkeson, A., and L.E. Ohanian (2001). Are Phillips curves useful for forecasting inflation?. Federal Reserve Bank of Minneapolis Quarterly Review 25, 2–11.

Bauwens, L., and M. Lubrano (1998). Bayesian inference on GARCH models using Gibbs sampler. Econometrics Journal 1, 23–46.

Bauwens, L., M. Lubrano, and J.-F. Richard (1999). Bayesian Inference in Dynamic Econometric Models, Oxford University Press.

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

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

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

Fama, E. F., and M. R. Gibbons (1984). A comparison of inflation forecasts. Journal of Monetary Economics 13, 327–348.

Galí, J., and M. Gertler, (1999). Inflation dynamics: A structural econometric analysis. Journal of Monetary Economics 44, 195–222.

Galí, J., M. Gertler, and J.D. López-Salidod (2005). Robustness of the estimates of the hybrid New Keynesian Phillips curve. Journal of Monetary Economics 52, 1107–1118.

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

Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin (2004). Bayesian Data Analysis, 2nd edition, Chapman & Hall/CRC.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernado, J.M., J.O. Berger, A.P. Dawid, and A.F.M. Smith (eds.) Bayesian Statistics 4, Clarendon Press, Oxford, UK.

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

Geweke, J. (1999). Using simulation methods for Bayesian econometric models: Inference, development and communication. Econometric Reviews 18, 1–126.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky (1999). Bayesian model averaging: A tutorial**.** Statistical Science 14, 382–417.

Huang, J., and Y. Pawitan (2000). Quasi-likelihood estimation of noninvertible moving average processes. Scandinavian Journal of Statistics 27, 689–710.

Koop, G. (2003). Bayesian Econometrics. Wiley.

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

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

Nelson, C.R., and G.W. Schwert (1977). Short-term interest rates as predictors of inflation: On testing the hypothesis that the real rate of interest is constant. American Economic Review 67, 478–486.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. Journal of the Royal Statistical Society Ser. B 57, 99–138.

Rudd, J., and K. Whelan (2006). Can rational expectations sticky-price models explain inflation dynamics? American Economic Review 96, 303–320.

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

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

Table 1. Simulation results for Student prior and multivariate Normal prior.

		DGP					
		$\rho_1 = 0.1$	$\phi_1 = 0.7$	$\rho_1 = 0.7$	$\phi_1 = 0.1$	$\rho_1 = 0.7$	$\phi_1 = 0.7$
	Parameter	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
Posterior mode Student prior	ρ_1	0.100	0.079	0.678	0.057	0.690	0.061
	ϕ_1	0.680	0.062	0.106	0.072	0.687	0.062
Posterior mode Normal prior	ρ_1	0.103	0.084	0.684	0.067	0.687	0.064
	ϕ_1	0.681	0.066	0.101	0.080	0.690	0.064
ML estimator	ρ_1	0.102	0.086	0.681	0.067	0.686	0.067
	ϕ_1	0.687	0.067	0.108	0.081	0.692	0.067

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 $p(M_4|\mathbf{y})$ and proportion of times when $p(M_4|\mathbf{y}) > 0.5$.

Panel A. The mean of $p(M_4|\mathbf{y})$

	ϕ_1					
ρ_1	0.1	0.3	0.5	0.7	0.9	0.99
0.0	0.18	0.29	0.30	0.25	0.20	0.11
0.1	0.26	0.45	0.49	0.45	0.39	0.24
0.3	0.44	0.84	0.91	0.95	0.95	0.77
0.5	0.45	0.92	0.99	1.00	1.00	0.99
0.7	0.40	0.93	1.00	1.00	1.00	1.00
0.9	0.27	0.89	1.00	1.00	1.00	1.00
0.99	0.15	0.76	0.99	1.00	1.00	1.00

Panel B. Proportion of times when $p(M_4|\mathbf{y}) > 0.5$

	ϕ_1					
ρ_1	0.1	0.3	0.5	0.7	0.9	0.99
0.0	0.06	0.12	0.13	0.09	0.07	0.06
0.1	0.11	0.37	0.44	0.36	0.33	0.22
0.3	0.36	0.89	0.93	0.97	0.97	0.77
0.5	0.41	0.95	0.99	1.00	1.00	0.98
0.7	0.35	0.96	1.00	1.00	1.00	1.00
0.9	0.23	0.91	1.00	1.00	1.00	1.00
0.99	0.10	0.77	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 $v = 3$ and $\sigma = 1$. The results are based on 1,000 realizations of a series of 150 observations.

Table 3. Mean of p(Noncausality|**y**) and proportion of times when p(Noncausality|**y**) > 0.5.

Panel A. The mean of the posterior probability $p(M_3|\mathbf{y})+p(M_4|\mathbf{y})$ of the noncausal process

	ϕ_1					
ρ_1	0.1	0.3	0.5	0.7	0.9	0.99
0.0	0.57	0.94	0.99	0.99	0.94	0.63
0.1	0.58	0.94	0.99	0.99	0.95	0.72
0.3	0.50	0.93	0.99	1.00	1.00	0.89
0.5	0.47	0.93	1.00	1.00	1.00	0.99
0.7	0.41	0.94	1.00	1.00	1.00	1.00
0.9	0.32	0.90	1.00	1.00	1.00	1.00
0.99	0.46	0.84	1.00	1.00	1.00	1.00

Panel B. Proportion of times when $p(M_3|\mathbf{y})+p(M_4|\mathbf{y}) > 0.5$

	ϕ_1					
ρ_1	0.1	0.3	0.5	0.7	0.9	0.99
0.0	0.54	0.98	0.99	1.00	0.96	0.66
0.1	0.59	0.97	1.00	1.00	0.98	0.79
0.3	0.50	0.96	1.00	1.00	1.00	0.91
0.5	0.44	0.96	1.00	1.00	1.00	0.99
0.7	0.36	0.96	1.00	1.00	1.00	1.00
0.9	0.28	0.92	1.00	1.00	1.00	1.00
0.99	0.44	0.87	1.00	1.00	1.00	1.00

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

Table 4. Estimation results of the equation (1) for the U.S. inflation

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

	s				
r	0	1	2	3	4
0	0.000	0.000	0.001	0.279	0.078
1	0.000	0.000	0.004	0.138	0.070
2	0.000	0.000	0.061	0.027	0.021
3	0.026	0.034	0.183	0.016	0.012
4	0.006	0.011	0.032	0.005	0.007
$p(s = \bullet \|\mathbf{y})$	0.032	0.045	0.281	0.465	0.188

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

Posterior/Par.	ϕ_1	ϕ_2	ϕ_3	ν	σ
Median	0.286	0.265	0.284	4.546	2.230
Std.	0.078	0.073	0.072	2.426	0.306

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

Table 5. The root mean squared forecast errors (RMSE) for the competitive models

	AR(r_{aic},0)	IMA(1,1)	Naïve	AR(r_m,0)	AR(r_m, s_m)	BMA
	The out-of-sample forecasting period, 1984:1 - 2006:4					
RMSE(h=1)	1.00	1.01	1.28	0.98	**0.97**	0.98
RMSE(h=2)	1.00	1.01	1.04	1.00	**0.97**	**0.97**
RMSE(h=3)	1.00	0.94	0.94	0.98	**0.92**	0.93
RMSE(h=4)	1.00	0.93	**0.87**	0.96	**0.87**	0.90
	The out-of-sample forecasting period, 1984:1 - 1994:2					
RMSE(h=1)	1.00	0.99	1.19	1.00	**0.98**	**0.98**
RMSE(h=2)	1.00	0.95	0.96	1.00	**0.94**	0.96
RMSE(h=3)	1.00	**0.91**	0.95	0.99	0.93	0.96
RMSE(h=4)	1.00	0.90	**0.87**	0.95	**0.87**	0.94
	The out-of-sample forecasting period, 1994:3 - 2006:4					
RMSE(h=1)	1.00	1.02	1.33	0.98	**0.97**	**0.97**
RMSE(h=2)	1.00	1.05	1.10	1.00	0.99	**0.97**
RMSE(h=3)	1.00	0.97	0.94	0.98	0.90	**0.89**
RMSE(h=4)	1.00	0.96	0.87	0.97	0.88	**0.86**

Reported figures are the RMSE of the forecasting models relative to the AR(AIC) benchmark model. Bold entries denote the lowest RMSE for period/horizon

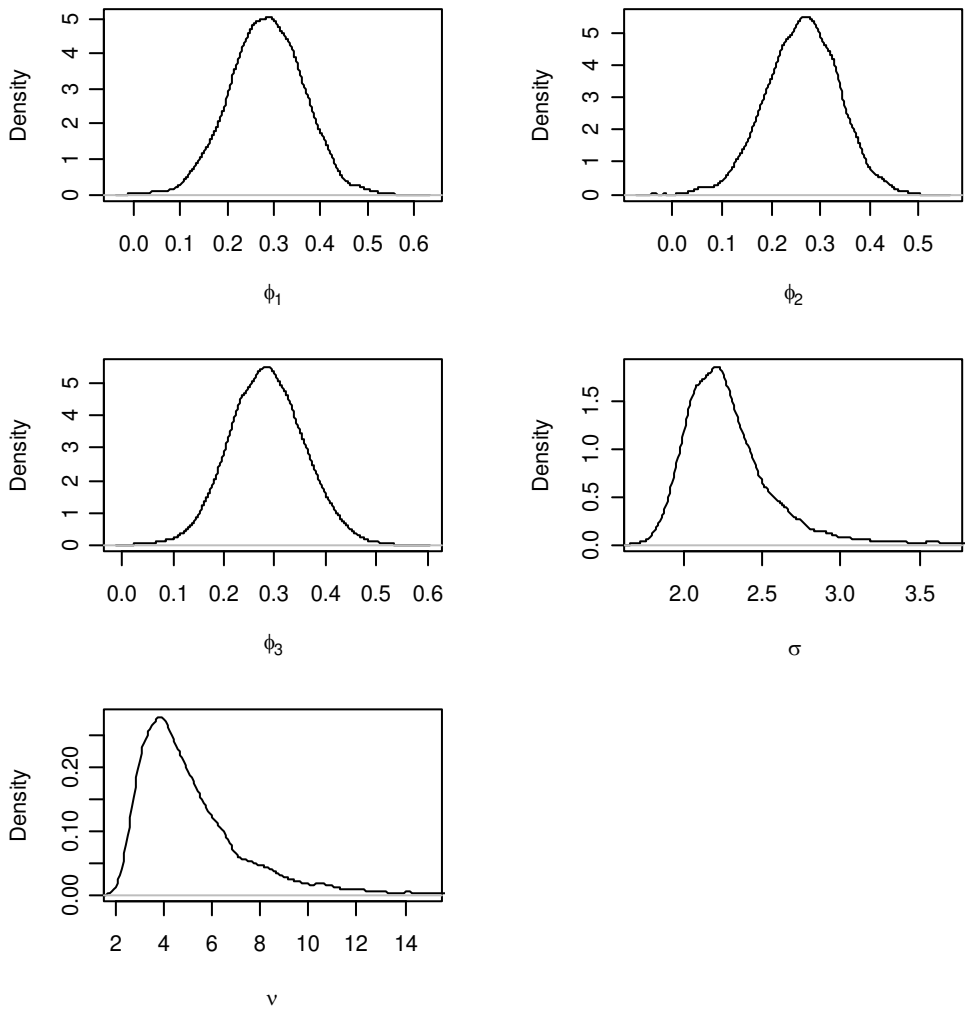

Figure 1. Posterior densities of parameters of the maximum a posteriori model for the U.S. inflation