



Helsinki  
Center  
of  
Economic  
Research

Discussion Papers

# Modeling Expectations with Noncausal Autoregressions

Markku Lanne  
University of Helsinki, RUESEG and HECER

and

Pentti Saikkonen  
University of Helsinki, RUESEG and HECER

Discussion Paper No. 212  
April 2008

ISSN 1795-0562

# Modeling Expectations with Noncausal Autoregressions\*

## Abstract

This paper is concerned with univariate noncausal autoregressive models and their potential usefulness in economic applications. We argue that noncausal autoregressive models are especially well suited for modeling expectations. Unlike conventional causal autoregressive models, they explicitly show how the considered economic variable is affected by expectations and how expectations are formed. Noncausal autoregressive models can also be used to examine the related issue of backward-looking or forward-looking dynamics of an economic variable. We show in the paper how the parameters of a noncausal autoregressive model can be estimated by the method of maximum likelihood and how related test procedures can be obtained. Because noncausal autoregressive models cannot be distinguished from conventional causal autoregressive models by second order properties or Gaussian likelihood, a detailed discussion on their specification is provided. Motivated by economic applications we explicitly use a forward-looking autoregressive polynomial in the formulation of the model. This is different from the practice used in previous statistics literature on noncausal autoregressions and, in addition to its economic motivation, it is also convenient from a statistical point of view. In particular, it facilitates obtaining likelihood based diagnostic tests for the specified orders of the backward-looking and forward-looking autoregressive polynomials. Such test procedures are not only useful in the specification of the model but also in testing economically interesting hypotheses such as whether the considered variable only exhibits forward-looking behavior. As an empirical application, we consider modeling the U.S. inflation dynamics which, according to our results, is purely forward-looking.

**JEL Classification:** C22, C52, E31

**Keywords:** Noncausal autoregression, expectations, inflation persistence

Markku Lanne

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

e-mail: [markku.lanne@helsinki.fi](mailto:markku.lanne@helsinki.fi)

Pentti Saikkonen

Department of Mathematics and Statistics  
P.O. Box 68 (Gustaf Hällströmn katu 2b)  
FI-00014 University of Helsinki  
FINLAND

e-mail: [pentti.saikkonen@helsinki.fi](mailto:pentti.saikkonen@helsinki.fi)

\* We thank Antti Ripatti for useful comments. Financial support from the Academy of Finland and the Okobank Group Research Foundation is gratefully acknowledged. This research was completed while the second author was a Fernand Braudel Fellow in the Economics Department of the European University Institute.

## 1 Introduction

Univariate autoregressive models are commonly employed in analyzing economic time series. Typical fields of application include forecasting and the measurement of persistence (Andrews and Chen (1994)), but the dynamics of state variables is also often modeled as an autoregressive process in macroeconomic (see, e.g., Canova (2007)) and financial (see, e.g., Campbell et al. (1997)) models. However, to the best of our knowledge, all economic applications so far restrict themselves to causal autoregressive models where the current value of the variable of interest is forced to depend only on its past. Noncausal autoregressive models, in contrast, also allow for dependence on the future. In our view, this is a particularly useful feature in economic applications where expectations play a central role. Indeed, noncausal autoregressive models conveniently facilitate explicitly modeling both backward-looking and forward-looking dynamics which has been of considerable interest in the recent macroeconomic literature (see, for instance, the literature on inflation persistence discussed in Section 5 below). Noncausal autoregressive models also lend themselves to a convenient economic interpretation. In particular, they make explicit how expectations of future error terms of the model affect both the current value and expected future values of the variable of interest.

In statistics literature, noncausal autoregressive and autoregressive moving average models have been studied, *inter alia*, by Breidt et al. (1991), Lii and Rosenblatt (1996), Huang and Pawitan (2000), Rosenblatt (2000), Breidt et al. (2001), and Andrews et al. (2006). However, this literature is not voluminous and, as discussed in these papers, typical applications have been confined to natural sciences and engineering.<sup>1</sup> In many of these applications it may actually not be reasonable to think of the

---

<sup>1</sup>As far as we know, the only empirical example of noncausal autoregressive moving average models with economic data is provided by Breidt et al. (2001) who demonstrate that a noncausal first order autoregressive model is appropriate for modeling a daily time series of Microsoft trading volume. Empirical economic examples of related models with a noninvertible moving average part are given in Huang and Pawitan (2000) and Breidt et al. (2001). In the former paper a noninvertible

employed model as a time series model but rather as a one-dimensional random field in which the direction of “time” is irrelevant and prediction is not of interest. This is in stark contrast with economics where the value added of the extension to the non-causal case most likely lies in the possibility of examining the effects of expectations of the future on the current value of an economic variable.

This paper demonstrates the potential that noncausal autoregressive models can have in economic applications. Unlike in the aforementioned previous literature, our formulation of the model explicitly involves a forward-looking autoregressive polynomial. This is in line with the practice of explicitly including expectations in economic models, and it also has statistical advantages. Indeed, a useful implication of our formulation is that statistical inference on autoregressive parameters is facilitated and it becomes, for example, straightforward to obtain likelihood based diagnostic tests for the specified orders of the backward-looking and forward-looking autoregressive polynomials. Obtaining specification tests of this kind within in the previously employed formulation appears less straightforward. A further advantage is that the autoregressive parameters are orthogonal to the parameters in the distribution of the error term so that inference on these two sets of parameters is asymptotically independent.

Once allowance for noncausality is made, model selection becomes a more complicated empirical issue than in conventional causal autoregressions. Which model is selected is also of great economic interest, as it tells us whether an economic variable exhibits backward-looking or forward-looking behavior or their combination. One well-known complication with noncausal autoregressions is that a non-Gaussian error term is required to achieve identification. In previous economic applications, causal autoregressive models with Gaussian error terms have typically been assumed. However, this approach has usually been justified by quasi maximum likelihood (ML) arguments because significant departures from Gaussianity, especially excess kurtosis. 

---

moving average model is applied to U.S. unemployment rate whereas the latter uses the so-called all-pass model to New Zealand/U.S. exchange rate. No discussion about expectations is provided in these papers, however.

sis, have been detected by diagnostic checks. In this paper, an error term with a t-distribution is found to provide an adequate fit but other leptokurtic distributions could also be considered. Once the distribution of the error term has been specified, we follow Breidt et al. (1991) and consider, in addition to diagnostic tests, a model selection algorithm based on the maximized log-likelihood function.

The proposed model is applied to study the U.S. inflation dynamics. A large part of the related previous literature based on univariate methods concentrates on the finding that inflation seems to be highly persistent which is considered to be in contrast with typical New Keynesian models assuming inflation to be forward-looking. Previous empirical results are based on conventional causal autoregressive models in which high persistence indeed necessarily implies backward-looking behavior. However, our results suggest that a purely noncausal autoregressive model is a far better description for U.S. inflation. This implies that the persistence previously found with univariate methods is not caused by dependence on past inflation but by predictability inherent in the noncausal autoregressive nature of the process. Hence, in a univariate framework the U.S. inflation seems to be forward-looking despite the strong persistence.

The rest of the paper is organized as follows. In Section 2 the noncausal autoregressive model is introduced and its properties are discussed. Section 3 considers (approximate) ML estimation and statistical inference in noncausal autoregressive models. In Section 4 a small-scale simulation study is conducted to examine the practical relevance of the asymptotic results presented in Section 3 as well as the aforementioned model selection procedure. Section 5 presents an empirical application to U.S. inflation. Finally, Section 6 concludes.

## 2 Model

### 2.1 Definition and basic properties

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

$$\varphi(B^{-1})\phi(B)y_t = \epsilon_t, \quad (1)$$

where  $\phi(B) = 1 - \phi_1 B - \dots - \phi_r B^r$ ,  $\varphi(B^{-1}) = 1 - \varphi_1 B^{-1} - \dots - \varphi_s B^{-s}$ , and  $\epsilon_t$  is a sequence of independent, identically distributed (continuous) random variables with mean zero and variance  $\sigma^2$  or, briefly,  $\epsilon_t \sim i.i.d.(0, \sigma^2)$ . Moreover,  $B$  is the usual backward shift operator, that is,  $B^k y_t = y_{t-k}$  ( $k = 0, \pm 1, \dots$ ), and the polynomials  $\phi(z)$  and  $\varphi(z)$  have their zeros outside the unit circle so that

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

If  $\varphi_j \neq 0$  for some  $j \in \{1, \dots, s\}$ , equation (1) defines a noncausal autoregression referred to as purely noncausal when  $\phi_1 = \dots = \phi_r = 0$ . The conventional causal autoregression is obtained when  $\varphi_1 = \dots = \varphi_s = 0$ . Then the former condition in (2) guarantees the stationarity of the model. In the general set up of equation (1) the same is true for the process  $u_t = \varphi(B^{-1})y_t$  which has the backward-looking moving average representation

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

where  $\alpha_0 = 1$  and the coefficients  $\alpha_j$  decay to zero at a geometric rate as  $j \rightarrow \infty$ . Similarly, the latter condition in (2) guarantees the stationarity of the purely noncausal process  $v_t = \phi(B)y_t$  and the validity of its forward-looking moving average representation

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

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

$$y_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j}, \quad (5)$$

where  $\psi_j$  is the coefficient of  $z^j$  in the Laurent series expansion of  $\phi(z)^{-1} \varphi(z^{-1})^{-1} \stackrel{def}{=} \psi(z)$ . Specifically, by condition (2),

$$\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$$

exists in some annulus  $b < |z| < b^{-1}$  with  $b < 1$  and reduces to the one-sided special cases obtained from (3) and (4) when  $y_t$  is causal and purely noncausal, respectively. The representation (5) implies that  $y_t$  is a stationary and ergodic process with finite second moments. We use the abbreviation  $\text{AR}(r, s)$  for the model defined by (1). In the causal case  $s = 0$ , the conventional abbreviation  $\text{AR}(r)$  is also used.

In the previous literature on noncausal autoregressions, it has been common to specify the model as

$$a(B)y_t = \varepsilon_t, \tag{6}$$

where  $a(B) = 1 - a_1 B - \dots - a_p B^p$  with  $a_p \neq 0$  and  $\varepsilon_t$  is an *i.i.d.* sequence with zero mean and finite variance (see, e.g., Breidt et al. (1991), Rosenblatt (2000) and the references therein). In this set up the relevant stationarity condition is  $a(z) \neq 0$ ,  $|z| \neq 1$ . When it holds  $y_t$  has a two-sided moving average representation similar to that in (5) (see Brockwell and Davis (1987, p. 88)). Moreover, when  $p = r + s$  and the number of zeros of  $a(z)$  outside (inside) the unit circle is  $r$  ( $s$ ), one can factor the polynomial  $a(z)$  as

$$a(z) = \varphi^*(z) \phi(z), \tag{7}$$

where  $\phi(z)$  is as in (1) and  $\varphi^*(z) = 1 - \varphi_1^* z - \dots - \varphi_s^* z^s$  has its zeros inside the unit circle, that is,  $\varphi^*(z) \neq 0$  for  $|z| \geq 1$ . Note that this particularly means that in the noncausal case  $s > 0$  the condition  $|\varphi_s^*| > 1$  holds.

The polynomial  $\varphi^*(z)$  can be expressed as

$$\begin{aligned} \varphi^*(z) &= -\varphi_s^* z^s \left( 1 + \frac{\varphi_{s-1}^*}{\varphi_s^*} z^{-1} + \dots + \frac{\varphi_1^*}{\varphi_s^*} z^{1-s} - \frac{1}{\varphi_s^*} z^{-s} \right) \\ &= -\varphi_s^* z^s \varphi(z^{-1}), \end{aligned}$$

where  $\varphi(z^{-1})$  is as in (1) so that  $\varphi_{s-j}^*/\varphi_s^* = -\varphi_j$  for  $j = 1, \dots, s-1$  and  $1/\varphi_s^* = \varphi_s$ . Because the zeros of  $\varphi^*(z)$  lie inside the unit circle those of  $\varphi(z)$  lie outside the unit

circle, as can be readily checked. Thus, the latter condition in (2) holds and model (1) can be obtained from (6) by defining  $\epsilon_t = -(1/\varphi_s^*)\epsilon_{t+s}$ . Similarly, if  $\varphi_s \neq 0$  is assumed in (1) the preceding reasoning can be reversed to obtain the specification (6) with  $\epsilon_t = -(1/\varphi_s)\epsilon_{t-s}$  and the coefficients of the polynomial  $\varphi^*(z)$  in (7) given by  $\varphi_j^* = -\varphi_j/\varphi_s$ ,  $j = 1, \dots, s-1$ , and  $\varphi_s^* = 1/\varphi_s$ . Thus, when  $\varphi_s \neq 0$  there is a one-to-one correspondence between the parameters in (1) and (6).<sup>2</sup>

A practical complication of noncausal autoregressive processes is that they cannot be identified by second order properties or Gaussian likelihood. This can be seen as follows. First, conclude from well-known results on linear filters that the spectral density function of the process  $y_t$  defined by (1) (or (6) and (7)) is given by  $\sigma^2/2\pi |\phi(e^{-i\omega})\varphi(e^{-i\omega})|^2$ . The same spectral density is obtained from a causal autoregressive process with lag polynomial  $\varphi(B)\phi(B)$  having its zeros outside the unit circle. These observations explain that  $y_t$  also has the causal representation

$$\varphi(B)\phi(B)y_t = \xi_t, \quad (8)$$

where the (stationary) innovation sequence  $\xi_t$  is uncorrelated but, in general, not independent with mean zero and variance  $\sigma^2$  (cf. Brockwell and Davis (1987, p. 124-125)). Thus, even if  $y_t$  is noncausal, its spectral density and, hence, autocovariance function cannot be distinguished from those of a causal autoregressive process. Thus, before applying a noncausal model it is advisable in practice to first fit an (adequate) causal autoregression to the observed series by standard least squares or Gaussian ML and check whether the residuals look non-Gaussian.

Unless otherwise stated, we shall henceforth assume that  $\epsilon_t$  is non-Gaussian and that its distribution has a (Lebesgue) density  $f_\sigma(x; \lambda) = \sigma^{-1}f(\sigma^{-1}x; \lambda)$  which depends on the parameter vector  $\lambda$  ( $d \times 1$ ) in addition to the scale parameter  $\sigma$  introduced earlier.

---

<sup>2</sup>This kind of reparameterization of model (6) is mentioned in Lii and Rosenblatt (1996, p. 17) in the context of a noncausal and noninvertible autoregressive moving average model. However, in that paper the model is not explicitly written as in (1) and the case  $\varphi_s = 0$  allowed in (1) is not discussed.

The formulation (1) appears more convenient than (6) and (7) when one needs to specify the (usually) unknown model orders  $r$  and  $s$ . Indeed, it turns out to be quite feasible to construct conventional likelihood based tests for hypotheses such as  $\phi_{r_0+1} = \dots = \phi_r = 0$  ( $r_0 < r$ ) and  $\varphi_{s_0+1} = \dots = \varphi_s = 0$  ( $s_0 < s$ ). For the latter hypothesis similar test procedures seem to be more difficult to obtain if the model is formulated as in (6) and (7) because  $|\varphi_s^*| > 1$  by assumption and because the logarithm of  $|\varphi_s^*|$  appears in the likelihood function (see Breidt et al. (1991)).<sup>3</sup> A further statistical convenience of the specification (1) is that the autoregressive parameters  $\phi = (\phi_1, \dots, \phi_r)$  and  $\varphi = (\varphi_1, \dots, \varphi_s)$  turn out to be orthogonal to the parameters  $\sigma^2$  and  $\lambda$  implying asymptotic independence of the corresponding ML estimators.<sup>4</sup>

Allowing for noncausality complicates predicting the process  $y_t$  which is pertinent in economic applications when expectations are studied. Let  $\mathcal{F}_t$  be the information set ( $\sigma$ -algebra) generated by  $\{y_t, y_{t-1}, \dots\}$  and let  $E_t(\cdot)$  be the corresponding conditional expectation operator. In the following discussion it is convenient to use the formulation (6) from which it is seen that the optimal (in mean square sense) one-step ahead predictor of  $y_{t+1}$  based on  $\mathcal{F}_t$  satisfies

$$E_t(y_{t+1}) = a_1 y_t + \dots + a_p y_{t-p+1} + E_t(\varepsilon_{t+1}). \quad (9)$$

If  $y_t$  is noncausal, the conditional expectation on the right hand side does not vanish because then  $\varepsilon_{t+1}$  ( $= -(1/\varphi_s) \varepsilon_{t+1-s}$ ) is not independent of  $\mathcal{F}_t$  (see (5)). Of course, the situation is similar when predictions for longer time horizons are considered. Thus, for optimal prediction knowledge of the distribution of the error process  $\varepsilon_t$  is required and, even if this knowledge is available, prediction is not easy because, in general, the prediction problem is nonlinear. Indeed, it is shown in Rosenblatt (2000, Corollary

---

<sup>3</sup>For statistical inference the previously mentioned condition  $a_p \neq 0$  is not needed, as the definition of the parameter space used in Lii and Rosenblatt (1996, p. 16) indicates.

<sup>4</sup>We use the notation  $x = (x_1, \dots, x_n)$  to introduce the  $n$ -dimensional vector  $x$  and its components. The same convention is also used when the components are vectors. In matrix calculations all vectors are interpreted as column vectors and a prime is used to signify the transpose of a vector or a matrix.

5.4.2) that if  $\varepsilon_t$  is non-Gaussian with finite  $(k + 1)$ st cumulant for some integer  $k \geq 2$  and if the zeros of  $\varphi^*(z)$  are simple then the optimal one-step ahead predictor is necessarily nonlinear. If  $\varepsilon_t$  is Gaussian so is  $y_t$  and the prediction problem is linear, but this is of little practical interest because then the possible noncausal nature of the process cannot be empirically revealed.

Even if the distribution of the error process  $\varepsilon_t$  is known the conditional expectations needed to compute optimal predictions may be unobtainable analytically. It is known, however, that even in the noncausal case the process  $y_t$  is  $p$ th order Markovian so that the conditional expectations  $E_t(y_{t+h})$  ( $h \geq 1$ ) are functions of  $y_t, \dots, y_{t-p+1}$  only (see Rosenblatt (2000, p. 90–93)). Thus, these functions can be estimated by simulating a long realization from the considered noncausal autoregression, as described in Breidt et al. (1991), and using available nonparametric estimation methods. This approach may be used to obtain predictions in practice but working out its feasibility and theoretical properties is outside the scope of this paper.

## 2.2 Economic interpretation

The noncausal autoregressive model considered in the previous section is economically appealing as a description of how economic agents form expectations and how realized values are affected by expectations. We first demonstrate that the model implies that the current value of the process,  $y_t$ , is affected by expected future errors. Using the definition of the process  $v_t$  and taking conditional expectation with respect to  $\mathcal{F}_t$  on both sides of equation (4) yields

$$y_t = \phi_1 y_{t-1} + \dots + \phi_r y_{t-r} + \sum_{j=0}^{\infty} \beta_j E_t(\epsilon_{t+j}). \quad (10)$$

In a causal model,  $\beta_j = 0$ ,  $j > 0$ , and the last term is just  $\epsilon_t$  implying that expected future errors have no effect on  $y_t$ . However, as our discussion on equation (9) shows, the last term is generally nonzero in a noncausal model, indicating the potential dependence of  $y_t$  on (an infinite number of) expected future errors. Note also that in a noncausal model  $E_t(\epsilon_t) \neq \epsilon_t$  because  $\epsilon_t$  depends on  $y_{t+j}$  ( $0 < j \leq s$ ) (see (1)).

The model also shows how expectations are affected by future errors. Leading (4) by one period and taking conditional expectations with respect to  $\mathcal{F}_t$  on both sides gives

$$E_t(y_{t+1}) = \phi_1 y_t + \cdots + \phi_r y_{t-r+1} + \sum_{j=0}^{\infty} \beta_j E_t(\epsilon_{t+1+j}). \quad (11)$$

In a purely causal model, future errors have no effect on the conditional expectation of  $y_{t+1}$  because  $\beta_j = 0$ ,  $j > 0$ , and  $E_t(\epsilon_{t+1}) = 0$ . However, as already discussed, the last term is different from zero in a noncausal model, indicating that the conditional expectation of future errors directly affects the conditional expectation of  $y_{t+1}$ . In economic applications, this can be interpreted as the predictable part of future errors having an effect on expectations. Note that this particularly means that, in the noncausal case, the errors  $\epsilon_t$  cannot be interpreted as unpredictable shocks similar to those appearing in economic applications of conventional causal models. In the following discussion, we emphasize this distinction between causal and noncausal models and call the errors impulses.

Assuming, for simplicity, that the process is purely noncausal we shall now provide a more detailed discussion on the dynamics of the process and economic agents' expectation formation. For concreteness (and anticipating our empirical application), suppose that  $y_t$  is inflation at time  $t$  and consider how it is affected by  $E_t(\epsilon_{t+h})$  ( $h \geq 1$ ), an expected inflation impulse at a future time point  $t+h$  that will be held fixed. From equation (10) it is seen that the impact of the expected inflation impulse  $E_t(\epsilon_{t+h})$  on the current inflation  $y_t$  is given by the value of the coefficient  $\beta_h$  whereas equation (11) shows its impact on  $E_t(y_{t+1})$ , the expected next period's inflation. The latter impact is given by the value of the coefficient  $\beta_{h-1}$ . Due to unexpected inflation shocks the realized inflation at time point  $t+1$ , that is,  $y_{t+1}$  differs from the expectation  $E_t(y_{t+1})$  economic agents have made at time  $t$ . However, despite this discrepancy inflation at time  $t+1$  is still determined by expected future inflation impulses, that is, variables in the current information set, as seen from equation (10) by leading the time index by one. At time point  $t+1$ , the impact of the expected inflation impulse

$E_{t+1}(\epsilon_{t+h})$  on the current inflation  $y_{t+1}$  is given by the value of the coefficient  $\beta_{h-1}$  whereas its impact on  $E_{t+1}(y_{t+2})$ , the expected next period's inflation, is given by the value of the coefficient  $\beta_{h-2}$ . In this way the process evolves until the time point  $t+h$  where the expected inflation impulse  $E_{t+h}(\epsilon_{t+h})$  affects the realized inflation  $y_{t+h}$  with coefficient equal to unity. Thereafter the inflation impulse  $\epsilon_{t+h}$  has no impact on future inflation  $y_{t+h+j}$  ( $j \geq 1$ ). This is illustrated in Figure 1, where the dotted arrows below the time line give the impact of  $E_{t+1}(\epsilon_{t+h})$  and  $E_t(\epsilon_{t+h})$  on expected future inflation and those above the time line give their impact on realized inflation at time points  $t$  and  $t+1$ .

If the process is not purely noncausal, that is, if lagged values of inflation are included, the expected and realized inflation are, in addition, affected by the expected inflation impulses through past inflation in the same way as in causal autoregressive models. In particular, in this case,  $E_{t+h+j}(\epsilon_{t+h})$  also has an impact on the future inflation  $y_{t+h+j}$  ( $j \geq 1$ ). Notice that from equation (1) it is seen that in the non-causal case  $\epsilon_{t+h}$  depends on  $y_{t+h+i}$ ,  $i = 1, \dots, s$ , (or at least on some of them) so that  $E_{t+h+j}(\epsilon_{t+h}) = \epsilon_{t+h}$  generally holds only when  $j \geq s$ .

The preceding discussion shows that expected future inflation impulses related to any fixed time point  $t+h$  change when the time point  $t+h$  is approached so that economic agents' expectations are based on forecast horizons that get shorter and shorter. This means that the impact of the expected future inflation impulses on realized inflation actually consists of two factors. When  $t$  refers to the current time point these factors are  $E_t(\epsilon_{t+h})$ , the size of the expected future inflation impulse, and  $\beta_h$ , the related coefficient.

Thus, the model provides an illuminating description about the dynamics of the inflation process and economic agents' expectation formation. In particular, the model shows how economic agents adjust their expectations to unexpected inflation shocks once they become observable so that the realized inflation is always determined by expected future inflation impulses and (possibly) past inflation. To further illustrate this point, notice that the last term on the right hand side of equation (11) can be

written as  $E_t(v_{t+1})$  (see (4)) and  $E_t(v_{t+1}) \neq v_{t+1}$  because  $v_{t+1} = \phi(B)y_{t+1}$  is not determined by the information available at time point  $t$ . The difference between  $E_t(v_{t+1})$  and  $v_{t+1}$  can be interpreted as the unexpected inflation shock economic agents face at time point  $t + 1$ . However, they fully adjust to this unexpected shock because  $E_{t+1}(v_{t+1}) = v_{t+1}$  so that inflation at time  $t + 1$  is determined by variables in the current information set. Note that even though  $v_{t+1} = \sum_{j=0}^{\infty} \beta_j \epsilon_{t+1+j}$  is perfectly predictable at time point  $t + 1$  this is not the case for  $\epsilon_{t+1+j}$  ( $j \geq 0$ ) because these variables are not determined by the information available at time point  $t + 1$  (see equation (1)).

### 3 Parameter estimation and statistical inference

#### 3.1 Approximate likelihood function

ML estimation of the parameters of a noncausal autoregression was studied by Breidt et al. (1991) by using the formulation based on equation (6). Even in this set up our model is slightly more general than theirs because we allow the distribution of the error term to depend on the additional parameter vector  $\lambda$ . This generalization has been considered by Andrews et al. (2006) in a related context and, following the arguments used in their paper, it can also be straightforwardly handled in our case. Thus, we shall assume that the density function  $f(x; \lambda)$  satisfies the regularity conditions of Andrews et al. (2006) which, among other things, require that  $f(x; \lambda)$  is twice continuously differentiable with respect to  $(x, \lambda)$ , non-Gaussian, and positive for all  $x \in \mathbb{R}$  and all permissible values of  $\lambda$ . The permissible parameter space of  $\lambda$ , denoted by  $\Lambda$ , is some subset of  $\mathbb{R}^d$  whereas the permissible space of the parameters  $\phi$ ,  $\varphi$  and  $\sigma$  is defined by the conditions in (2) and by  $\sigma > 0$ . For convenience, the regularity conditions of Andrews et al. (2006) are also presented in the appendix and, unless otherwise stated, they will henceforth be assumed. Densities that satisfy these conditions include a rescaled  $t$ -density and a weighted average of Gaussian densities.

If the model is defined as in (6) and (7), ML estimators of the parameters in (1)

can be derived by a smooth one-to-one transformation from ML estimators of the parameters in (6), and hence their limiting distribution can also be easily obtained. However, because this reasoning is not directly applicable if the degree of the polynomial  $\varphi(z)$  is overspecified (i.e.,  $\varphi_s = 0$ ) we shall provide details based directly on the specification (1). We start by deriving the likelihood function.

Suppose we have an observed time series  $y_1, \dots, y_T$ . Using the definitions  $u_t = \varphi(B^{-1})y_t$  and  $v_t = \phi(B)y_t$  we can write

$$\begin{bmatrix} u_1 \\ \vdots \\ u_{T-s} \\ v_{T-s+1} \\ \vdots \\ v_T \end{bmatrix} = \begin{bmatrix} y_1 - \varphi_1 y_2 - \dots - \varphi_s y_{s+1} \\ \vdots \\ y_{T-s} - \varphi_1 y_{T-s+1} - \dots - \varphi_s y_T \\ y_{T-s+1} - \phi_1 y_{T-s} - \dots - \phi_r y_{T-s+1-r} \\ \vdots \\ y_T - \phi_1 y_{T-1} - \dots - \phi_r y_{T-r} \end{bmatrix} = A \begin{bmatrix} y_1 \\ \vdots \\ y_{T-s} \\ y_{T-s+1} \\ \vdots \\ y_T \end{bmatrix}$$

or briefly

$$x = Ay.$$

Similarly,

$$\begin{bmatrix} u_1 \\ \vdots \\ u_r \\ \epsilon_{r+1} \\ \vdots \\ \epsilon_{T-s} \\ v_{T-s+1} \\ \vdots \\ v_T \end{bmatrix} = \begin{bmatrix} u_1 \\ \vdots \\ u_r \\ u_{r+1} - \phi_1 u_r - \dots - \phi_r u_1 \\ \vdots \\ u_{T-s} - \phi_1 u_{T-s-1} - \dots - \phi_r u_{T-s-r} \\ v_{T-s+1} \\ \vdots \\ v_T \end{bmatrix} = B \begin{bmatrix} u_1 \\ \vdots \\ u_r \\ u_{r+1} \\ \vdots \\ u_{T-s} \\ v_{T-s+1} \\ \vdots \\ v_T \end{bmatrix}$$

or

$$z = Bx.$$

Hence, the vectors  $z$  and  $y$  are related by

$$z = BAy.$$

Note that from (3) and (4) it can be seen that the components of  $z$  given by  $(u_1, \dots, u_r)$ ,  $(\epsilon_{r+1}, \dots, \epsilon_{T-s})$ , and  $(v_{T-s+1}, \dots, v_T)$  are independent. The joint density function of  $z$  under true parameter values can thus be expressed as

$$h_U(u_1, \dots, u_r) \left( \prod_{t=r+1}^{T-s} f_\sigma(\epsilon_t; \lambda) \right) h_V(v_{T-s+1}, \dots, v_T),$$

where  $h_U$  and  $h_V$  signify the joint density functions of  $(u_1, \dots, u_r)$  and  $(v_{T-s+1}, \dots, v_T)$ , respectively. It is easy to see that the (nonstochastic) matrices  $A$  and  $B$  are non-singular and the determinant of  $B$  is unity so that we can express the joint density function of the data vector  $y$  as

$$h_U(\varphi(B^{-1})y_1, \dots, \varphi(B^{-1})y_r) \left( \prod_{t=r+1}^{T-s} f_\sigma(\varphi(B^{-1})\phi(B)y_t; \lambda) \right) \\ \times h_V(\phi(B)y_{T-s+1}, \dots, \phi(B)y_T) \det(A).$$

It is also easy to check that the determinant of the  $(T-s) \times (T-s)$  block in the upper left hand corner of  $A$  is unity and, using the well-known formula for the determinant of a partitioned matrix, it can furthermore be seen that the determinant of  $A$  is independent of the sample size  $T$ . This suggests approximating the joint density of  $y$  by the second factor in the preceding expression, giving rise to the approximate log-likelihood function

$$l_T(\theta) = \sum_{t=r+1}^{T-s} g_t(\theta), \quad (12)$$

where  $\theta = (\phi, \varphi, \sigma, \lambda)$  and

$$g_t(\theta) = \log f(\sigma^{-1}(u_t(\varphi) - \phi_1 u_{t-1}(\varphi) - \dots - \phi_r u_{t-r}(\varphi)); \lambda) - \log \sigma \\ = \log f(\sigma^{-1}(v_t(\phi) - \varphi_1 v_{t+1}(\phi) - \dots - \varphi_s v_{t+s}(\phi)); \lambda) - \log \sigma.$$

Here  $u_t(\varphi)$  and  $v_t(\phi)$  signify the series  $u_t$  and  $v_t$  treated as functions of the parameters  $\varphi$  and  $\phi$ , respectively. Maximizing  $l_T(\theta)$  over permissible values of  $\theta$  gives an

approximate ML estimator of  $\theta$ . Note that here, as well as in the next section, the orders  $r$  and  $s$  are assumed known. Procedures to specify these quantities will be discussed in later sections of the paper.

### 3.2 Asymptotic properties of the approximate ML estimator

In what follows, it will be convenient to use the notation  $\theta_0$  for the true value of  $\theta$  and similarly for its components. It is assumed that  $\lambda_0$ , the true value  $\lambda$ , is an interior point of  $\Lambda$ .

We shall first consider the score of  $\theta$  evaluated at true parameter values. Define the vectors  $U_{t-1} = (u_{t-1}, \dots, u_{t-r})$  and  $V_{t+1} = (v_{t+1}, \dots, v_{t+s})$  where  $u_t$  and  $v_t$  are defined in terms of true parameter values so that  $u_t = \sum_{j=0}^{\infty} \alpha_{0j} \epsilon_{t-j}$  and  $v_t = \sum_{j=0}^{\infty} \beta_{0j} \epsilon_{t+j}$ . By straightforward differentiation (cf. Breidt et al. (1991)) we find from (12) that

$$\frac{\partial}{\partial \phi} g_t(\theta_0) = -\frac{f'(\sigma_0^{-1} \epsilon_t; \lambda_0)}{\sigma_0 f(\sigma_0^{-1} \epsilon_t; \lambda_0)} U_{t-1} \quad (r \times 1)$$

and

$$\frac{\partial}{\partial \varphi} g_t(\theta_0) = -\frac{f'(\sigma_0^{-1} \epsilon_t; \lambda_0)}{\sigma_0 f(\sigma_0^{-1} \epsilon_t; \lambda_0)} V_{t+1} \quad (s \times 1),$$

where  $f'(x, \lambda) = \partial f(x, \lambda) / \partial x$  and use has also been made of the fact that  $\phi_0(B) u_t = \epsilon_t = \varphi_0(B) v_t$  with  $\phi_0(B)$  and  $\varphi_0(B)$  defined in terms of true parameter values (e.g.  $\phi_0(B) = 1 - \phi_{01}B - \dots - \phi_{0r}B^r$ ). Similarly,

$$\frac{\partial}{\partial \sigma} g_t(\theta_0) = -\sigma_0^{-2} \left( \frac{f'(\sigma_0^{-1} \epsilon_t; \lambda_0)}{f(\sigma_0^{-1} \epsilon_t; \lambda_0)} \epsilon_t + \sigma_0 \right)$$

and

$$\frac{\partial}{\partial \lambda} g_t(\theta_0) = \frac{1}{f(\sigma_0^{-1} \epsilon_t; \lambda_0)} \frac{\partial}{\partial \lambda} f(\sigma_0^{-1} \epsilon_t; \lambda_0) \quad (d \times 1).$$

The following lemma presents the asymptotic distribution of the score vector. For the presentation of this lemma we need some notation. Let  $\eta_t \sim i.i.d.(0, 1)$  and define the AR( $r$ ) process  $u_t^*$  by  $\phi_0(B) u_t^* = \eta_t$  and the AR( $s$ ) process  $v_t^*$  by  $\varphi_0(B) v_t^* = \eta_t$ . Note that  $u_t^*$  and  $v_t^*$  are jointly stationary and causal with finite second moments. Next form the vectors  $U_{t-1}^* = (u_{t-1}^*, \dots, u_{t-r}^*)$  and  $V_{t-1}^* = (v_{t-1}^*, \dots, v_{t-s}^*)$

and the associated covariance matrices  $\Gamma_{U^*} = Cov(U_{t-1}^*)$ ,  $\Gamma_{V^*} = Cov(V_{t-1}^*)$ , and  $\Gamma_{U^*V^*} = Cov(U_{t-1}^*, V_{t-1}^*) = \Gamma_{V^*U^*}$ . We also define

$$\mathcal{J} = \int \frac{(f'(x; \lambda_0))^2}{f(x; \lambda_0)} dx$$

and set

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} \mathcal{J}\Gamma_{U^*} & \Gamma_{U^*V^*} \\ \Gamma_{V^*U^*} & \mathcal{J}\Gamma_{V^*} \end{bmatrix}.$$

Note that  $\Gamma_{U^*} = \sigma_0^{-2}Cov(U_{t-1})$ ,  $\Gamma_{V^*} = \sigma_0^{-2}Cov(V_{t+1})$ , and  $\mathcal{J} > 1$  (see condition (A5) of Andrews et al. (2006)). Finally, define the  $(d+1) \times (d+1)$  matrix

$$\Omega = \begin{bmatrix} \omega_\sigma^2 & \omega_{\sigma\lambda} \\ \omega_{\lambda\sigma} & \Omega_{\lambda\lambda} \end{bmatrix}, \quad (13)$$

where

$$\Omega_{\lambda\lambda} = \int \frac{1}{f(x; \lambda_0)} \left( \frac{\partial}{\partial \lambda} f(x; \lambda_0) \right) \left( \frac{\partial}{\partial \lambda} f(x; \lambda_0) \right)' dx,$$

$$\omega_{\lambda\sigma} = -\sigma_0^{-1} \int x \frac{f'(x; \lambda_0)}{f(x; \lambda_0)} \frac{\partial}{\partial \lambda} f(x; \lambda_0) dx = \omega'_{\sigma\lambda},$$

and

$$\omega_\sigma^2 = \sigma_0^{-2} \left( \int x^2 \frac{(f'(x; \lambda_0))^2}{f(x; \lambda_0)} dx - 1 \right).$$

Now we can present the limiting distribution of the score vector.<sup>5</sup>

**Lemma 1** *If conditions (A1)–(A7) of Andrews et al. (2006) hold, then*

$$(T-p)^{-1/2} \sum_{t=r+1}^{T-s} \frac{\partial}{\partial \theta} g_t(\theta_0) \xrightarrow{d} N(0, \text{diag}(\Sigma, \Omega)).$$

Moreover, the matrices  $\Sigma$  and  $\Omega$  are positive definite.

Lemma 1 can be proved in the same way as Propositions 1 and 2 of Breidt et al. (1991). An outline of the needed arguments is provided in the appendix. Here we note that the positive definiteness of the matrix  $\Sigma$  follows from the above mentioned

---

<sup>5</sup>The notation  $\text{diag}(A_1, A_2)$  signifies a block diagonal matrix with diagonal blocks  $A_1$  and  $A_2$ .

inequality  $\mathcal{J} > 1$  which holds when  $\epsilon_t$  is non-Gaussian (see Remark 2 of Andrews et al. (2006)). The matrix  $\Sigma$  is positive definite even if the model order  $r$  or  $s$  is overspecified or both are overspecified. For instance, suppose that  $r = s$  and consider the extreme case where  $\phi = \varphi = 0$ . Then,  $\Sigma_{11} = \Sigma_{22} = \mathcal{J}I_r$  and  $\Sigma_{12} = I_r$  so that the matrix  $\Sigma$  is clearly positive definite when  $\mathcal{J} > 1$ . In the general case of Lemma 1 the positive definiteness of the matrix  $\Omega$  must be assumed (cf. condition (A6) of Andrews et al. (2006)). The block diagonality of the covariance matrix of the limiting distribution implies that the scores of  $(\phi, \varphi)$  and  $(\sigma, \lambda)$  are asymptotically independent. This property, commonly referred to as orthogonality of the parameters  $(\phi, \varphi)$  and  $(\sigma, \lambda)$ , is convenient because it means that statistical inference on the autoregressive parameters  $\phi$  and  $\varphi$ , which is typically of primary interest, is asymptotically independent of the estimation of the parameters  $\sigma$  and  $\lambda$  describing the distribution of the error term  $\epsilon_t$ . It may be noted that similar orthogonality does not hold if the formulation given by (6) and (7) is used because then the score of the autoregressive parameter  $\varphi_s^*$  is asymptotically correlated with the score of the scale parameter of the error term  $\epsilon_t$  (see Proposition 2 of Breidt et al. (1991)).

Using a conventional Taylor series expansion of the score in conjunction with Lemma 1 and the assumed regularity conditions one can show the existence of a consistent and asymptotically normal (local) maximizer of the approximate likelihood function. Specifically, the following theorem can be established. Its proof makes use of arguments similar to those in Breidt et al. (1991) and Andrews et al. (2006) and is outlined in the appendix.

**Theorem 2** *If conditions (A1)–(A7) of Andrews et al. (2006) hold, there exists a sequence of (local) maximizers  $\hat{\theta} = (\hat{\phi}, \hat{\varphi}, \hat{\sigma}, \hat{\lambda})$  of  $l_T(\theta)$  in (12) such that*

$$(T - p)^{1/2} (\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \text{diag}(\Sigma^{-1}, \Omega^{-1})).$$

Due to the block diagonality of the covariance matrix of the limiting distribution, the (approximate) ML estimators  $(\hat{\phi}, \hat{\varphi})$  and  $(\hat{\sigma}, \hat{\lambda})$  are asymptotically independent.

This means that if a consistent initial estimator  $(\tilde{\phi}, \tilde{\varphi})$  of  $(\phi, \varphi)$  is available an estimator of  $(\sigma, \lambda)$  with the same asymptotic distribution as the ML estimator  $(\hat{\sigma}, \hat{\lambda})$  can be obtained by maximizing the function  $l_T(\tilde{\phi}, \tilde{\varphi}, \sigma, \lambda)$ . As the initial estimator  $(\tilde{\phi}, \tilde{\varphi})$  one may consider the least absolute deviation (LAD) estimator based on the (possibly incorrect) assumption that  $\epsilon_t$  has a Laplace (or double exponential) distribution. In the case of the specification (6) Huang and Pawitan (2000) establish the consistency of the LAD estimator when, in a certain sense, the true distribution of  $\epsilon_t$  has tails heavier than the normal distribution. Their result applies to a variety of known distributions including the  $t$ -distribution and normal scale mixtures. An inspection of the residuals based on a LAD estimation may also help to specify an appropriate distribution for the error term  $\epsilon_t$ .

### 3.3 Statistical inference

To be able to compute approximate standard errors for the components of the estimator  $\hat{\theta}$  and construct confidence intervals and conventional Wald tests we need consistent estimators of the covariance matrices  $\Sigma$  and  $\Omega$ . First consider the former and define the  $p \times 1$  vector  $W_t^* = (U_t^*, V_t^*)$ . Here  $U_t^*$  ( $r \times 1$ ) and  $V_t^*$  ( $s \times 1$ ) are as in the previous section so that their components are defined in terms of the stationary AR( $r$ ) and AR( $s$ ) processes  $\phi_0(B)u_t^* = \eta_t$  and  $\varphi_0(B)v_t^* = \eta_t$ , respectively. Consider the equation

$$W_t^* = DW_{t-1}^* + \iota\eta_t,$$

where the 1st and  $(r+1)$ th components of the vector  $\iota$  ( $(r+s) \times 1$ ) are unity and the remaining components are zero,  $D = \text{diag}(D_\phi, D_\varphi)$  is a block diagonal matrix with

$$D_\phi = \begin{bmatrix} \phi_{01} & \phi_{02} & \cdots & \phi_{0,r-1} & \phi_{0r} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \quad (r \times r),$$

and  $D_\varphi$  ( $s \times s$ ) defined in the same way but using  $\varphi_{01}, \dots, \varphi_{0s}$  instead of  $\phi_{01}, \dots, \phi_{0r}$ . From the definitions it follows that  $E\eta_t W_{t-1}^* = 0$  so that, because  $\eta_t \sim i.i.d. (0, 1)$ , we get the equation

$$EW_t^* W_t^{*'} = DEW_{t-1}^* W_{t-1}^{*'} D' + \iota'. \quad (14)$$

The process  $W_t^*$  is covariance stationary and, therefore,  $EW_t^* W_t^{*'} = EW_{t-1}^* W_{t-1}^{*'} = \Gamma_{W^*}$ , say. From the definitions it can be seen that

$$\Gamma_{W^*} = \begin{bmatrix} \Gamma_{U^*} & \Gamma_{U^* V^*} \\ \Gamma_{V^* U^*} & \Gamma_{V^*} \end{bmatrix},$$

which differs from  $\Sigma$  only in that the scalar factor  $\mathcal{J}$  has been omitted from the diagonal blocks. Thus, once we have consistent estimators for  $\Gamma_{W^*}$  and  $\mathcal{J}$  we immediately obtain a consistent estimator for  $\Sigma$ .

Let  $vec$  denote the usual vectorization operator which stacks columns of a matrix one below another. From equation (14) and well-known properties of the  $vec$  operator one obtains

$$vec(\Gamma_{W^*}) = (D \otimes D) vec(\Gamma_{W^*}) + \iota \otimes \iota,$$

where  $\otimes$  signifies the Kronecker product. By stationarity, the eigenvalues of the matrix  $D$  are inside the unit circle so that the same is true for the eigenvalues of  $D \otimes D$ . From the preceding equation one thus obtains

$$vec(\Gamma_{W^*}) = \left( I_{(r+s)^2} - D \otimes D \right)^{-1} (\iota \otimes \iota).$$

Replacing  $D$  on the right hand side with the obvious ML estimator  $\hat{D}$  one obtains a consistent estimator for  $vec(\Gamma_{W^*})$  from which a consistent estimator for  $\Gamma_{W^*}$  is obtained. This estimator is denoted by  $\hat{\Gamma}_{W^*}$  with a similar notation used for its blocks (e.g.  $\hat{\Gamma}_{U^*}$ ).

We still need a consistent estimator for  $\mathcal{J}$ . The definition of  $\mathcal{J}$  suggests the obvious estimator

$$\hat{\mathcal{J}} = \int \frac{(f'(x; \hat{\lambda}))^2}{f(x; \hat{\lambda})} dx$$

the consistency of which follows from that of  $\hat{\lambda}$  and the regularity conditions assumed (see Remark 7 of Andrews et al. (2007)). Thus, a consistent estimator of  $\Sigma$  is given by

$$\hat{\Sigma} = \begin{bmatrix} \hat{\mathcal{J}}\hat{\Gamma}_{U^*} & \hat{\Gamma}_{U^*V^*} \\ \hat{\Gamma}_{V^*U^*} & \hat{\mathcal{J}}\hat{\Gamma}_{V^*} \end{bmatrix}.$$

In the special case of a causal ( $s = 0$ ) or a purely noncausal model ( $r = 0$ ) the estimation of the covariance matrix  $\Sigma$  simplifies because then  $\Sigma$  reduces to  $\Sigma_{11}$  or  $\Sigma_{22}$ , respectively. In particular cases it may also be easy to obtain an explicit expression for  $\mathcal{J}$ . For instance, if  $f_\sigma(x; \lambda) = \sigma^{-1}f(\sigma^{-1}x; \lambda)$  is a rescaled  $t$ -density with  $\lambda$  degrees of freedom we have

$$\mathcal{J} = \frac{\lambda_0(\lambda_0 + 1)}{(\lambda_0 - 2)(\lambda_0 + 3)}$$

(cf. Remark 4 of Andrews et al. (2006) and the definition of the quantity  $\tilde{J}$  in Theorem 1 of that paper). Thus, a consistent estimator of  $\mathcal{J}$  can be obtained by replacing  $\lambda_0$  on the right hand side with its ML estimator  $\hat{\lambda}$ .

Consistent estimators of the components of the matrix  $\Omega$  can be constructed as in Remark 7 of Andrews et al. (2006). Specifically, one can use the estimators

$$\hat{\Omega}_{\lambda\lambda} = \int \frac{1}{f(x; \hat{\lambda})} \left( \frac{\partial}{\partial \lambda}(fx; \hat{\lambda}) \right) \left( \frac{\partial}{\partial \lambda}(fx; \hat{\lambda}) \right)' dx,$$

$$\hat{\omega}_{\lambda\sigma} = -\hat{\sigma}^{-1} \int x \frac{f'(x; \hat{\lambda})}{f(x; \hat{\lambda})} \frac{\partial}{\partial \lambda} f(x; \hat{\lambda}) dx,$$

and

$$\hat{\omega}_\sigma^2 = \hat{\sigma}^{-2} \left( \int x \frac{(f'(x; \hat{\lambda}))^2}{f(x; \hat{\lambda})} dx - 1 \right).$$

In particular cases it may be possible to replace the integrals by more explicit expressions, as in the case of the quantity  $\mathcal{J}$  when the rescaled  $t$ -distribution was assumed.

In general, one may also make use of the orthogonality of the parameters  $(\phi, \varphi)$  and  $(\sigma, \lambda)$  and obtain a consistent estimator of  $\Omega$  by computing the value of the matrix  $-(T - p)^{-1} \partial^2 l_T(\hat{\phi}, \hat{\varphi}, \theta_2) / \partial \theta_2 \partial \theta_2'$  at  $\hat{\theta}_2 = (\hat{\sigma}, \hat{\lambda})$  numerically.

Approximate standard errors of the components of  $\hat{\theta}$  can be obtained by computing the square roots of the diagonal elements of the matrices  $(T - p)^{-1} \hat{\Sigma}$  and  $(T - p)^{-1} \hat{\Omega}$ .

Conventional Wald tests are also readily obtained. As an example, consider testing the null hypotheses

$$H : \phi_{0,r_0+1} = \cdots = \phi_{0r} = 0 \quad \text{and} \quad \varphi_{0,s_0+1} = \cdots = \varphi_{0s} = 0,$$

where  $r_0 < r$  and  $s_0 < s$ . Thus, the null hypothesis implies that the model orders  $r$  and  $s$  can be reduced to  $r_0$  and  $s_0$ , respectively, with the case  $r_0 = r$  or  $s_0 = s$  obtained with obvious modification. To present the test, define the  $(r - r_0) \times r$  matrix  $R_\phi = [0 : I_{r-r_0}]$  and the  $(s - s_0) \times s$  matrix  $R_\varphi = [0 : I_{s-s_0}]$  and set  $R = \text{diag}(R_\phi, R_\varphi)$ . Then the null hypothesis can be expressed as  $R\theta_{01} = 0$ , where  $\theta_{01} = (\phi_0, \varphi_0) ((r + s) \times 1)$ . Denoting the ML estimator of  $\theta_1$  by  $\hat{\theta}_1$  we can write the conventional Wald test statistic as

$$\mathcal{W} = (T - p) \hat{\theta}'_1 R' (R \hat{\Sigma}^{-1} R')^{-1} R \hat{\theta}_1 \xrightarrow{d} \chi_{r-r_0+s-s_0},$$

where the convergence assumes the null hypothesis and is an immediate consequence of Theorem 2. Of course, the same result is obtained for any choice the matrix  $R$  with full row rank.

One may also use the likelihood ratio (LR) test. Let  $\tilde{\theta}$  signify the ML estimator of the parameter  $\theta$  constrained by the null hypothesis so that  $\tilde{\theta}$  is obtained by applying ML in the model with orders  $r_0$  and  $s_0$ . The LR test statistic is

$$\mathcal{LR} = l_T(\hat{\theta}) - l_T(\tilde{\theta}) \xrightarrow{d} \chi_{r-r_0+s-s_0},$$

where the null hypothesis is again assumed. The limiting distribution can be justified by a standard application of the results given in the appendix which can also be used to obtain the corresponding score (Lagrange multiplier) test. To the best of our knowledge, test procedures of this kind have not been explicitly considered in the previous literature of noncausal autoregressive models where the model is formulated as in (6) and (7) and treating the case  $s_0 < s$  is hampered by the condition  $|\varphi_s^*| > 1$ .

## 4 Simulation study

To study the finite-sample properties of the estimators and tests proposed in Section 3, we conducted a small simulation study. Following Breidt et al. (1991), we concentrate on the second-order process as the data-generating process (DGP) because it is the simplest model that allows for a versatile analysis of various aspects of estimation and testing. Throughout, the results are based on 10,000 realizations. We generate each realization in two steps. First, a series from the causal AR( $r$ ) model  $\phi(B)v_t = \epsilon_t$  ( $t = r+1, \dots, T$ ) is generated. Then  $y_t$  is computed recursively from  $\varphi(B^{-1})y_t = v_t$  for  $t = T-s, \dots, 1$ . The  $r$  and  $s$  initial values, respectively, are set to zero, and to eliminate initialization effects 100 observations at the beginning and end of each realization are discarded. In all experiments, the error term  $\epsilon_t$  is assumed to follow the  $t$ -distribution with 3 degrees of freedom and  $\sigma$  is set equal to 0.1. We consider three different combinations of parameter values,  $(\phi_1, \varphi_1) = \{(0.9, 0.9), (0.9, 0.1), (0.1, 0.9)\}$ . In the first case, the roots of the lag polynomials are equal and close to the unit circle, in the two other cases the roots of the “causal” and “noncausal” polynomials are clearly different. Three sample sizes, 100, 200 and 500 are considered.

The mean and standard deviation of the ML estimators of  $\phi_1$  and  $\varphi_1$  are presented in Table 1. Even with as few as 100 observations the parameters are relatively accurately estimated in each case, and the biases as well as the standard deviations clearly diminish as the sample size increases. In the case  $(\phi_1, \varphi_1) = (0.9, 0.9)$ ,  $\phi_1$  is more accurately estimated in terms of both criteria, whereas in the other two cases it is the parameter taking the smaller value that is estimated with a somewhat smaller bias. The differences are, however, minor.

The results concerning the Wald and LR tests of hypotheses involving a single parameter in Table 2 indicate that the Wald test tends to overreject even in samples as large as 500 observations. Although the overrejection problem is mitigated with the sample size, the rejection rates exceed 6% in a test with 5% nominal significance level still with 1,000 observations (not reported). The LR test, on the other hand, has

reasonable size properties in samples with 200 observations or more. For the Wald test, the case  $(\phi_1, \varphi_1) = (0.9, 0.9)$  seems to be the most difficult, while there are only minor differences in the rejection rates of the LR test across the different parameter values. In general, the tests on the parameter with the smaller value have somewhat better size properties, in accordance with the properties of the ML estimator above. Based on these results, the LR test can be recommended instead of the Wald test.

As the Wald test tends to overreject, we only present simulation results on power for the LR test. Because the size properties do not differ much between the different DGP's, only the rejection rates of the LR test (at the nominal 5% level of significance) for the first DGP  $((\phi_1, \varphi_1) = (0.9, 0.9))$  are presented in Figure 2. Moreover, we concentrate on tests concerning  $\phi_1$  because there is no reason to expect the power properties to greatly depend on the particular parameter. The values of  $\phi_1$  in the alternative DGP's that are used to generate the data are given by  $0.9 - c/\sqrt{T}$  ( $c = 0.0, 0.2, 0.4, \dots, 2.0$ ), and the null hypothesis in the test is  $\phi_1 = 0.9$ . The rejection rates for alternatives very close to the null are moderate for all sample sizes considered ( $T = 100, 200, 500$ ), but they rapidly increase with  $c$ , especially with the greater sample sizes. Hence, the LR test seems to have reasonable power. These results, however, suggest that in small samples, one should not rely on this test alone in model selection.

Breidt et al. (1991) suggested a model selection procedure based on maximizing the likelihood function. In other words, all purely causal, noncausal and mixed models of a given order ( $p$ ) are estimated, and the model yielding the greatest value of the likelihood function is selected. Their simulation results lend support to this procedure, and in Table 3, we present similar results when the DGP is the mixed second-order model. The procedure seems to work relatively well even with 100 observations, and the performance greatly improves with the sample size. However, there seem to be some differences depending on the parameter values. When  $(\phi_1, \varphi_1) = (0.9, 0.9)$ , the correct model is selected in 95% of the realizations with 200 observations, and the corresponding figure is 99.9% with 500 observations. In contrast, in the cases with

different parameter values, the causal (noncausal) model is selected far too often when  $\phi_1$  is smaller (greater), even with 500 observations. In these cases model selection is presumably complicated by the fact that the considered processes are rather close to first-order processes. Although the proposed procedure works fairly well even in these difficult cases, additional simulation experiments involving greater values of the other parameter (not reported) indicated improvements, with the correct model sometimes being selected even more frequently than in the  $(\phi_1, \varphi_1) = (0.9, 0.9)$  case. Despite the quite satisfactory performance of this procedure, the results suggest that model selection should not be based on this criterion alone, but, in addition, diagnostic tests should be employed.

## 5 Empirical application

In this section, we apply the models and methods discussed above to modeling U.S. inflation dynamics. Our focus is on examining the nature of inflation persistence that has given rise to a voluminous literature in the past few decades. The central question in this line of research is whether inflation is a purely forward-looking variable as required by typical New Keynesian models. This assumption has been tested by checking for serial correlation in inflation, and typically measures based on univariate autoregressive models such as the cumulative impulse response (CIR) (Andrews and Chen (1994)), have indicated quite high persistence of inflation in industrialized countries (for a survey of recent empirical literature, see Cecchetti and Debelle (2006)). The presence of high autocorrelation has been interpreted as evidence against the forward-looking inflation expectations assumed in the New Keynesian models. This, in turn, has led to modifications of existing theory that try to explain the apparently backward-looking behavior. This paper contributes to the large empirical literature that studies inflation persistence in the univariate framework only. To relate inflation persistence explicitly to macroeconomic theories of price determination, remaining persistence after taking the effect of variables such as marginal costs or output gap

into account should rather be considered. This extension of our model is, however, outside the scope of this paper.

To the best of our knowledge, only causal autoregressive models have been entertained in the previous literature on inflation. As a consequence, high persistence has automatically been interpreted as evidence of a large backward-looking component in inflation (see Cecchetti and Debelle (2006), and the references therein). However, as discussed in Section 2, high autocorrelation and, hence, strong persistence do not, per se, indicate strong backward-looking behavior. Even completely forward-looking inflation may be persistent if autocorrelation is used as a measure of persistence. The same is true if the CIR based on a causal autoregressive model is used to measure persistence. Indeed, as seen in Section 2, for any purely noncausal autoregressive process there is a corresponding causal process with the same lag polynomial and, hence, the same autocorrelation function and impulse response function. Thus, causality or noncausality and, hence, backward-looking or forward-looking dynamics, cannot be distinguished by examining the autocorrelation function or the impulse response function of a causal autoregressive model fitted to the series.

In what follows, we will use the procedures proposed earlier in the paper to argue that the U.S. inflation series is purely noncausal despite its strong persistence. This can be interpreted as evidence in favor of the forward-looking behavior of typical New Keynesian models. Hence, the strong autocorrelation and large CIR are not brought about by dependence on past inflation but by expectations of future errors to inflation.

The inflation series that we model, is the annualized quarterly inflation rate computed from the seasonally adjusted U.S. consumer price index (for all urban consumers) published by the Bureau of Labor Statistics. The sample period comprises 148 observations, from 1970:1 to 2006:4. There is positive autocorrelation even at high lags as shown by the autocorrelation function depicted in Figure 3. The Ljung-Box test indicates that autocorrelation is also significant at all reasonable significance levels. However, by visual inspection and unit root tests, the series can be consid-

ered stationary. Further evidence of persistence is provided by the CIR based on the causal Gaussian AR(3) model that turned out to adequately capture the linear dependence in the inflation series (see Table 4). The CIR of this model equals 5.8, which is comparable to the values obtained by Cecchetti and Debelle (2006) for the OECD countries, indicating high persistence.

In Table 4, we present the estimation results of a number of autoregressive models for the demeaned inflation, along with some diagnostic tests.<sup>6</sup> Of Gaussian autoregressive models up to order 4, the AR(3) model (AR(3,0)-N) was selected by both the Akaike (AIC) and Bayesian (BIC) information criteria. However, the diagnostic tests suggest that this model is misspecified. Although the Ljung-Box test does not indicate the presence of unmodeled autocorrelation, there is evidence of conditional heteroskedasticity, as the p-value of the McLeod-Li test is 0.004.<sup>7</sup> Moreover, the quantile-quantile plot of the residuals in the upper panel of Figure 4 indicates that the normal distribution fails to capture the tails of the error distribution. Also, normality of the quantile residuals of the AR(3,0)-N model is clearly rejected by the Shapiro-Wilk test (p-value is 0.001). These findings suggest that a more leptokurtic distribution, such as the  $t$ -distribution with a relatively small degrees-of-freedom parameter might provide a more satisfactory fit.

Because a Gaussian AR(3) model is deemed adequate in describing the autocorrelation structure of the inflation series, we proceed by estimating all alternative causal and noncausal AR( $r, s$ ) models with  $r + s = 3$ , following the procedure proposed in

---

<sup>6</sup>Estimation is done using the BHHH algorithm in the GAUSS CMLMT library.

<sup>7</sup>Note that, when the orders of the model are misspecified, the Ljung-Box and McLeod-Li tests are not exactly valid as they do not take estimation errors correctly into account. The reason is that a misspecification of the model orders makes the errors dependent, as pointed out in the case of the causal specification (8). Nevertheless, p-values of these tests can be seen as convenient summary measures of the autocorrelation remaining in residuals and their squares. A similar remark applies to the Shapiro-Wilk test used for quantile-quantile plots.

Section 4. The error term is assumed to have a  $t$ -distribution.<sup>8</sup> Of the four models, the purely noncausal model (AR(0,3)- $t$ ) maximizes the log-likelihood function by a clear margin to the other specifications. Furthermore, according to the diagnostic tests, this is the only specification that does not suffer from remaining autocorrelation and conditional heteroskedasticity. Interestingly, the evidence of remaining conditional heteroskedasticity diminishes as the importance of forward-looking dynamics increases, but of the noncausal models it is only in the purely forward-looking AR(0,3)- $t$  model that the error term does not exhibit autocorrelation. The adequacy of this model was also checked by testing it against higher-order specifications, and the coefficients of the additional terms turned out to be insignificant in the LR test (the p-values of an extra parameter in AR(1,3)- $t$  and AR(0,4)- $t$  models, are 0.08 and 0.14, respectively). Hence, the results attest to purely forward-looking inflation dynamics, indicating that it is the expectations of future errors that drive the inflation process.

In all cases, the degrees-of-freedom parameter  $\lambda$  is estimated very small, indicating fat-tailed error distributions. This is not surprising given the bad fit of the Gaussian AR(3) model. The quantile-quantile plot of the AR(0,3)- $t$  model depicted in the lower panel of Figure 4 lends support to the adequacy of the  $t$ -distribution, as does the Shapiro-Wilk test with p-value 0.45. In particular, the tail area seems to be better captured than under the normality assumption. As a matter of fact, all models with  $t$ -distributed errors generated a similar quantile-quantile plot, indicating that great improvements in fit are brought about by only properly selecting the error distribution.

---

<sup>8</sup>The log-likelihood function equals

$$l_T(\theta) = \sum_{t=r+1}^{T-s} g_t(\theta),$$

where

$$g_t(\theta) = \log \left\{ \frac{\Gamma[(\lambda+1)/2]}{\pi^{1/2}\Gamma(\lambda/2)} (\lambda-2)^{-1/2} \left[ 1 + \frac{\sigma^{-2}\epsilon_t^2}{\lambda-2} \right]^{-(\lambda+1)/2} \right\} - \log \sigma.$$

In summary, the results strongly indicate purely forward-looking inflation dynamics. Hence, the apparent persistence in inflation observed in univariate analyses is not caused by relying on past inflation in forming expectations but by the predictability of future inflation impulses to inflation. As a matter of fact, if the inflation series follows a noncausal autoregressive process, the true persistence of a shock to the inflation series may be different from that implied by the autocorrelation function or the CIR based on a causal autoregressive model. Moreover, because optimal predictions in the noncausal autoregressive model are nonlinear, persistence and the shape of the impulse response function may depend on the sign and size of a shock as well as the initial values. While the computation of the CIR is straightforward in the case of a causal autoregressive model, it becomes difficult when noncausality is present. In this case, tracing the effects of a shock calls for computing conditional expectations which, as pointed out in Section 2, are not available in closed form but require simulation methods. This issue lies outside the scope of this paper.

## 6 Conclusion

In this paper, we have considered univariate noncausal autoregressive models that, to the best of our knowledge, have so far not attracted attention in the economics and finance literature. In the applications presented in the related statistics literature, the direction of time has typically been an irrelevant aspect which is not the case in economic applications where expectations of the future play a central role. Therefore, we argue that allowing for noncausality opens up new possibilities for modeling expectations and their effects on the dynamics of economic variables. In particular, these models facilitate determining whether expectations are forward-looking as commonly assumed in theoretical economic models.

We discuss ML estimation and develop related tests for noncausal autoregressive models. Furthermore, based on a number of simulation experiments and our experience with actual economic data, we propose the following procedure for specifying

a potentially noncausal autoregressive model. The first step is to fit a conventional causal autoregressive model by least squares or Gaussian ML and determine its order by using conventional procedures such as diagnostic checks and model selection criteria. Once an adequate causal model is found, its error term should be tested for Gaussianity. Because identification requires the error term be non-Gaussian, we can proceed only if deviations from Gaussianity are detected. A variety of error distributions can be considered; in our empirical application we successfully employed the  $t$ -distribution. With the chosen error distribution, all causal and noncausal autoregressive models of the selected order are then estimated and the model maximizing the log-likelihood function is selected. Finally, through diagnostic tests the adequacy of this model is confirmed. These diagnostic checks should give information on directions in which the model potentially fails.

In future work, we plan to look at extensions of the univariate model considered in this paper. Being able to handle multiple times series is of interest, as our discussion about inflation persistence at the beginning of Section 5 indicates. Using noncausal autoregressions to model financial returns is another obvious field of application. To be able to adequately capture the erratic behavior of these time series probably calls for extensions of the basic model considered in this paper. In particular, allowing for forward-looking dynamics is hardly sufficient to model the conditional heteroskedasticity prevalent in financial returns.

## Mathematical appendix

We shall first present the regularity conditions (A1)–(A7) of Andrews et al. (2006). We use  $\Lambda_0 \subset \Lambda$  to signify some neighborhood of  $\lambda_0$ .

**(A1)** For all  $x \in \mathbb{R}$  and all  $\lambda \in \Lambda$ ,  $f(x; \lambda) > 0$  and  $f(x; \lambda)$  is twice continuously differentiable with respect to  $(x, \lambda)$ .

**(A2)** For all  $\lambda \in \Lambda_0$ ,  $\int x f'(x; \lambda) dx = x f(x; \lambda) \Big|_{-\infty}^{\infty} - \int f(x; \lambda) dx = -1$ .

**(A3)**  $\int f''(x; \lambda_0) dx = f'(x; \lambda_0) \Big|_{-\infty}^{\infty} = 0$ .

**(A4)**  $\int x^2 f''(x; \lambda_0) dx = x^2 f'(x; \lambda_0) \Big|_{-\infty}^{\infty} - 2 \int x f'(x; \lambda_0) dx = 2$ .

**(A5)**  $1 < \int (f'(x; \lambda_0))^2 / f(x; \lambda_0) dx$ .

**(A6)** The matrix  $\Omega$  defined in (13) is positive definite.

**(A7)** For  $j, k = 1, \dots, d$  and all  $\lambda \in \Lambda_0$ ,

- $f(x; \lambda)$  is dominated by a function  $f_1(x)$  such that  $\int x^2 f_1(x) dx < \infty$ , and
- $x^2 \frac{(f'(x; \lambda))^2}{f(x; \lambda)}$ ,  $x^2 \left| \frac{f''(x; \lambda)}{f(x; \lambda)} \right|$ ,  $|x| \left| \frac{\partial f'(x; \lambda) / \partial \lambda_j}{f(x; \lambda)} \right|$ ,  $\frac{(\partial f'(x; \lambda) / \partial \lambda_j)^2}{f^2(x; \lambda)}$ , and  $\frac{|\partial^2 f(x; \lambda) / \partial \lambda_j \partial \lambda_k|}{f(x; \lambda)}$  are dominated by  $a_1 + a_2 |x|^{c_1}$ , where  $a_1$ ,  $a_2$ , and  $c_1$  are nonnegative constants and  $\int |x|^{c_1} f_1(x) dx < \infty$ .

**Proof of Lemma 1.** First consider the covariance matrix of the score. For simplicity, denote  $e_t = f'(\sigma_0^{-1} \epsilon_t; \lambda_0) / [f(\sigma_0^{-1} \epsilon_t; \lambda_0) \sigma_0] = f'_{\sigma_0}(\sigma_0^{-1} \epsilon_t; \lambda_0) / f_{\sigma_0}(\sigma_0^{-1} \epsilon_t; \lambda_0)$  and notice that

$$\begin{aligned} E(e_t^2) &= E \left[ \left( f'_{\sigma_0}(\epsilon_t; \lambda_0) / f_{\sigma_0}(\epsilon_t; \lambda_0) \right)^2 \right] \\ &= \sigma_0^{-2} \int (f'(x; \lambda_0))^2 / f(x; \lambda_0) dx \\ &= \sigma_0^{-2} \mathcal{J}, \end{aligned}$$

where the second equality is based on the fact that  $f_{\sigma_0}(x; \lambda_0) = \sigma_0^{-1} f(\sigma_0^{-1}x; \lambda_0)$  is the density function of  $\epsilon_t$  (cf. equation (2.13) of Breidt et al. (1991)). Thus, because  $e_t$  and  $U_{t-1}$  are independent and  $\Gamma_{U^*} = \sigma_0^{-2} \text{Cov}(U_{t-1})$ ,

$$\begin{aligned} \text{Cov}\left(\frac{\partial}{\partial\phi}g_t(\theta_0)\right) &= \text{Cov}(-U_{t-1}e_t) \\ &= E(e_t^2)\text{Cov}(U_{t-1}) \\ &= \mathcal{J}\Gamma_{U^*}. \end{aligned}$$

Because the sequence  $U_{t-1}e_t$  is uncorrelated we have

$$\lim_{T \rightarrow \infty} (T-p)^{-1} \text{Cov}\left(\sum_{t=r+1}^{T-s} \frac{\partial}{\partial\phi}g_t(\theta_0), \sum_{t=r+1}^{T-s} \frac{\partial}{\partial\phi}g_t(\theta_0)\right) = \mathcal{J}\Gamma_{U^*}.$$

Similarly, the independence of  $e_t$  and  $V_{t+1}$  and the equality  $\Gamma_{V^*} = \sigma_0^{-2} \text{Cov}(V_{t+1})$  give

$$\text{Cov}\left(\frac{\partial}{\partial\varphi}g_t(\theta_0)\right) = \mathcal{J}\Gamma_{V^*}$$

and, by the uncorrelatedness of the sequence  $V_{t+1}e_t$ ,

$$\lim_{T \rightarrow \infty} (T-p)^{-1} \text{Cov}\left(\sum_{t=r+1}^{T-s} \frac{\partial}{\partial\varphi}g_t(\theta_0), \sum_{t=r+1}^{T-s} \frac{\partial}{\partial\varphi}g_t(\theta_0)\right) = \mathcal{J}\Gamma_{V^*}.$$

As for the covariance matrix between  $\partial g_t(\theta_0)/\partial\phi$  and  $\partial g_t(\theta_0)/\partial\varphi$ , first consider

$$\begin{aligned} \text{Cov}(-u_{t-i}e_t, -v_{k+j}e_k) &= \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \alpha_{0a}\beta_{0b} \text{Cov}(\epsilon_{t-i-a}e_t, \epsilon_{k+j+b}e_k) \\ &= \begin{cases} \alpha_{0,t-k-i}\beta_{0,t-k-j}, & t > k, 1 \leq i \leq r, 1 \leq j \leq s \\ 0, & t \leq k, 1 \leq i \leq r, 1 \leq j \leq s \end{cases}, \end{aligned}$$

where the first equality follows from (3) and (4) and the second one is based on condition (A2) (see also Breidt et al. (1991, p. 181)). Hence, as in Breidt et al. (1991, p. 182), the element in position  $(i, j)$  of the matrix  $(T-p)^{-1} \text{Cov}(\partial l_T(\theta_0)/\partial\phi, \partial l_T(\theta_0)/\partial\varphi)$  is

$$\begin{aligned} (T-p)^{-1} \sum_{k=r+1}^{T-s-1} \sum_{t=k+1}^{T-s} \alpha_{0,t-k-i}\beta_{0,t-k-j} &= (T-p)^{-1} \sum_{k=r+1}^{T-s-1} \sum_{t=0}^{T-s-k-i} \alpha_{0t}\beta_{0,t+i-j} \\ &\rightarrow \sum_{k=0}^{\infty} \alpha_{0k}\beta_{0,k+i-j}, \end{aligned}$$

where  $\beta_{0l} = 0$  for  $l < 0$ . Note that the limit equals  $\psi_{0,j-i}$ , as can be easily checked.

Next recall that  $u_t^* = \sum_{k=0}^{\infty} \alpha_{0k} \eta_{t-k}$  and  $v_t^* = \sum_{l=0}^{\infty} \beta_{0l} \eta_{t-l}$  with  $\eta_t \sim i.i.d. (0, 1)$ .

Thus,

$$\begin{aligned} Cov(u_{t-i}^*, v_{t-j}^*) &= \sum_{k=0}^{\infty} \alpha_{0k} \sum_{l=0}^{\infty} \beta_{0l} E(\eta_{t-i-k} \eta_{t-j-l}) \\ &= \sum_{k=0}^{\infty} \alpha_{0k} \beta_{0,k+i-j}, \end{aligned}$$

and we can conclude that

$$\lim_{T \rightarrow \infty} (T-p)^{-1} Cov \left( \sum_{t=r+1}^{T-s} \frac{\partial}{\partial \phi} g_t(\theta_0), \sum_{t=r+1}^{T-s} \frac{\partial}{\partial \varphi} g_t(\theta_0) \right) = \Gamma_{U^*V^*}.$$

We have thus shown that the covariance matrix of the score of  $(\phi, \varphi)$  evaluated at the true parameter value and divided by  $(T-p)$  converges to  $\Sigma$ .

The score of  $(\sigma, \lambda)$  is *i.i.d.* and, by condition (A7), has zero mean and finite second moments. The definitions show that its covariance matrix equals that of the score of the parameter  $(\alpha_{p+1}, \theta)$  in Andrews et al. (2006). Thus, if  $\theta_2 = (\sigma, \lambda)$

$$(T-p)^{-1} Cov \left( \sum_{t=r+1}^{T-s} \frac{\partial}{\partial \theta_2} g_t(\theta_0), \sum_{t=r+1}^{T-s} \frac{\partial}{\partial \theta_2} g_t(\theta_0) \right) = \Omega.$$

Using the definitions it is also straightforward to check that, at true parameter values, the scores of  $(\phi, \varphi)$  and  $(\sigma, \lambda)$  are uncorrelated so that we can conclude that

$$\lim_{T \rightarrow \infty} (T-p)^{-1} Cov \left( \sum_{t=r+1}^{T-s} \frac{\partial}{\partial \theta} g_t(\theta_0), \sum_{t=r+1}^{T-s} \frac{\partial}{\partial \theta_2} g_t(\theta_0) \right) = \text{diag}(\Sigma, \Omega).$$

The matrix  $\Omega$  is positive definite by the assumed condition (A6). Because  $\mathcal{J} > 1$  (see condition (A5)) the positive definiteness of  $\Sigma$  can be established in the same way as Proposition 1 of Breidt et al. (1991).

The asymptotic normality can be proved in the same way as Proposition 2 of Breidt et al. (1991) by approximating the processes  $U_{t-1}$  and  $V_{t+1}$  by long moving averages and using a standard central limit theorem for finitely dependent stationary processes.

**Proof of Theorem 2.** We shall first present the second partial derivatives of the function  $g_t(\theta)$ . To simplify notation, we write  $\tilde{u}_t = u_t(\varphi)$  and  $\tilde{v}_t = v_t(\phi)$  and, furthermore,  $\tilde{U}_{t-1} = (\tilde{u}_{t-1}, \dots, \tilde{u}_{t-r})$  and  $\tilde{V}_{t+1} = (\tilde{v}_{t+1}, \dots, \tilde{v}_{t+s})$ . Similarly,  $\tilde{\epsilon}_t = \tilde{u}_t - \phi_1 \tilde{u}_{t-1} - \dots - \phi_r \tilde{u}_{t-r} = \tilde{v}_t - \varphi_1 \tilde{v}_{t+1} - \dots - \varphi_s \tilde{v}_{t+s}$  will signify  $\epsilon_t$  evaluated at an arbitrary point in the permissible parameter space, not the true parameter value. We also set  $h(x; \lambda) = f'(x; \lambda) / f(x; \lambda)$ , so that

$$h'(x; \lambda) = \frac{f''(x; \lambda)}{f(x; \lambda)} - \left( \frac{f'(x; \lambda)}{f(x; \lambda)} \right)^2,$$

and let  $Y_t$  stand for the  $r \times s$  matrix with elements  $y_{t-i+j}$  ( $i = 1, \dots, r, j = 1, \dots, s$ ). By straightforward differentiation (cf. Breidt et al. (1991), p. 187),

$$\begin{aligned} \partial^2 g_t(\theta) / \partial \phi \partial \phi' &= \sigma^{-2} h'(\sigma^{-1} \tilde{\epsilon}_t; \lambda) \tilde{U}_{t-1} \tilde{U}'_{t-1} \\ \partial^2 g_t(\theta) / \partial \varphi \partial \varphi' &= \sigma^{-2} h'(\sigma^{-1} \tilde{\epsilon}_t; \lambda) \tilde{V}_{t+1} \tilde{V}'_{t+1} \\ \partial^2 g_t(\theta) / \partial \sigma^2 &= 2\sigma^{-3} h(\sigma^{-1} \tilde{\epsilon}_t; \lambda) \tilde{\epsilon}_t + \sigma^{-4} h'(\sigma^{-1} \tilde{\epsilon}_t; \lambda) \tilde{\epsilon}_t^2 + \sigma^{-2} \\ \partial^2 g_t(\theta) / \partial \lambda \partial \lambda' &= \frac{1}{f(\sigma^{-1} \tilde{\epsilon}_t; \lambda)} \partial^2 f(\sigma^{-1} \tilde{\epsilon}_t; \lambda) / \partial \lambda \partial \lambda' \\ &\quad - \frac{1}{f^2(\sigma^{-1} \tilde{\epsilon}_t; \lambda)} (\partial f(\sigma^{-1} \tilde{\epsilon}_t; \lambda) / \partial \lambda) (\partial f(\sigma^{-1} \tilde{\epsilon}_t; \lambda) / \partial \lambda)' \\ \partial^2 g_t(\theta) / \partial \phi \partial \varphi' &= \sigma^{-2} h'(\sigma^{-1} \tilde{\epsilon}_t; \lambda) \tilde{U}_{t-1} \tilde{V}'_{t+1} + \sigma^{-1} h(\sigma^{-1} \tilde{\epsilon}_t; \lambda) Y_t \\ \partial^2 g_t(\theta) / \partial \phi \partial \sigma &= \sigma^{-3} h'(\sigma^{-1} \tilde{\epsilon}_t; \lambda) \tilde{\epsilon}_t \tilde{U}_{t-1} + \sigma^{-2} h(\sigma^{-1} \tilde{\epsilon}_t; \lambda) \tilde{U}_{t-1} \\ \partial^2 g_t(\theta) / \partial \phi \partial \lambda' &= -\sigma^{-1} \tilde{U}_{t-1} \partial h(\sigma^{-1} \tilde{\epsilon}_t; \lambda) / \partial \lambda' \\ \partial^2 g_t(\theta) / \partial \varphi \partial \sigma &= \sigma^{-3} h'(\sigma^{-1} \tilde{\epsilon}_t; \lambda) \tilde{\epsilon}_t \tilde{V}_{t+1} + \sigma^{-2} h(\sigma^{-1} \tilde{\epsilon}_t; \lambda) \tilde{V}_{t+1} \\ \partial^2 g_t(\theta) / \partial \varphi \partial \lambda' &= -\sigma^{-1} \tilde{V}_{t+1} \partial h(\sigma^{-1} \tilde{\epsilon}_t; \lambda) / \partial \lambda' \\ \partial^2 g_t(\theta) / \partial \sigma \partial \lambda' &= -\sigma^{-2} \tilde{\epsilon}_t \partial h(\sigma^{-1} \tilde{\epsilon}_t; \lambda) / \partial \lambda'. \end{aligned}$$

Using conditions (A2)–(A4) and calculations similar to those in Breidt et al. (1991, p. 181) it is not difficult to check that  $E[\partial^2 g_t(\theta_0) / \partial \theta \partial \theta'] = -diag(\Sigma, \Omega)$ .

As in Andrews et al. (2006), we now use the Taylor series expansion

$$\begin{aligned} \sum_{t=r+1}^{T-s} [g_t(\theta_0 + T^{-1/2}c) - g_t(\theta_0)] &= T^{-1/2} \sum_{t=r+1}^{T-s} c' \frac{\partial g_t(\theta_0)}{\partial \theta} + \frac{1}{2} T^{-1} \sum_{t=r+1}^{T-s} c' \frac{\partial^2 g_t(\theta_0)}{\partial \theta \partial \theta'} c \\ &\quad + \frac{1}{2} T^{-1} \sum_{t=r+1}^{T-s} c' \left( \frac{\partial^2 g_t(\theta_T^*(c))}{\partial \theta \partial \theta'} - \frac{\partial^2 g_t(\theta_0)}{\partial \theta \partial \theta'} \right) c, \end{aligned}$$

where  $c \in \mathbb{R}^{r+s+1+d}$  and the argument  $\theta_T^*(c)$  in the matrix of second partial derivatives means that each row is evaluated at an intermediate point lying between  $\theta_0$  and  $T^{-1/2}c$ . Thus, if  $\|\cdot\|$  signifies the Euclidean norm we have  $\sup_{c \in K} \|\theta_T^*(c) - \theta_0\| \rightarrow 0$  for any compact set  $K \subset \mathbb{R}^{r+s+1+d}$ . Moreover, using the dominance conditions in (A7) and arguments similar to those in Breidt et al. (1991, p. 186-190) it can be shown that a uniform law of large numbers for stationary ergodic processes applies to  $\partial^2 g_t(\theta)/\partial \theta \partial \theta'$  over any small enough compact neighborhood  $\theta_0$  (see Theorem A.2.2 in White (1994)). Thus, we can conclude that

$$T^{-1} \sum_{t=r+1}^{T-s} c' \left( \frac{\partial^2 g_t(\theta_T^*(c))}{\partial \theta \partial \theta'} - \frac{\partial^2 g_t(\theta_0)}{\partial \theta \partial \theta'} \right) c \xrightarrow{p} 0$$

for  $c$  belonging to any compact subset of  $\mathbb{R}^{r+s+1+d}$ . The proof can now be completed in the same way as the proof of Theorem 1 of Andrews et al. (2006).

## 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.
- Andrews, D. and W. Chen (1994). Approximately median-unbiased estimation of autoregressive models. *Journal of Business and Economic Statistics* 12, 187–204.
- 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. *The Annals of Statistics* 29, 919-946.
- Brockwell, P.J. and R.A. Davis (1987). *Time Series: Theory and Methods*. Springer-Verlag. New York.
- Campbell, J.Y., A.W. Lo, and A.C. MacKinlay (1997). *Econometrics of Financial Markets*. Princeton University Press. Princeton.
- Canova, F. (2007). *Methods for Applied Macroeconomic Research*. Princeton University Press. Princeton.
- Cecchetti, S.G. and G. Debelle (2006). Has the inflation process changed? *Economic Policy*, April 2006, 311–352.
- Huang, J. and Y. Pawitan (2000). Quasi-likelihood estimation of noninvertible moving average processes. *Scandinavian Journal of Statistics* 27, 689-710.
- Lii, K.-S. and M. Rosenblatt (1996). Maximum likelihood estimation for non-Gaussian nonminimum phase ARMA sequences. *Statistica Sinica* 6, 1-22.

Rosenblatt, M. (2000). Gaussian and Non-Gaussian Linear Time Series and Random Fields. Springer-Verlag, New York.

White, H. (1994). Estimation, Inference and Specification Analysis. Cambridge University Press. New York.

Figure 1: The impact of  $\epsilon_{t+h}$  on realized inflation (above the time line) and expected future inflation (below the time line) in periods  $t$  and  $t + 1$  in a purely noncausal autoregressive model.

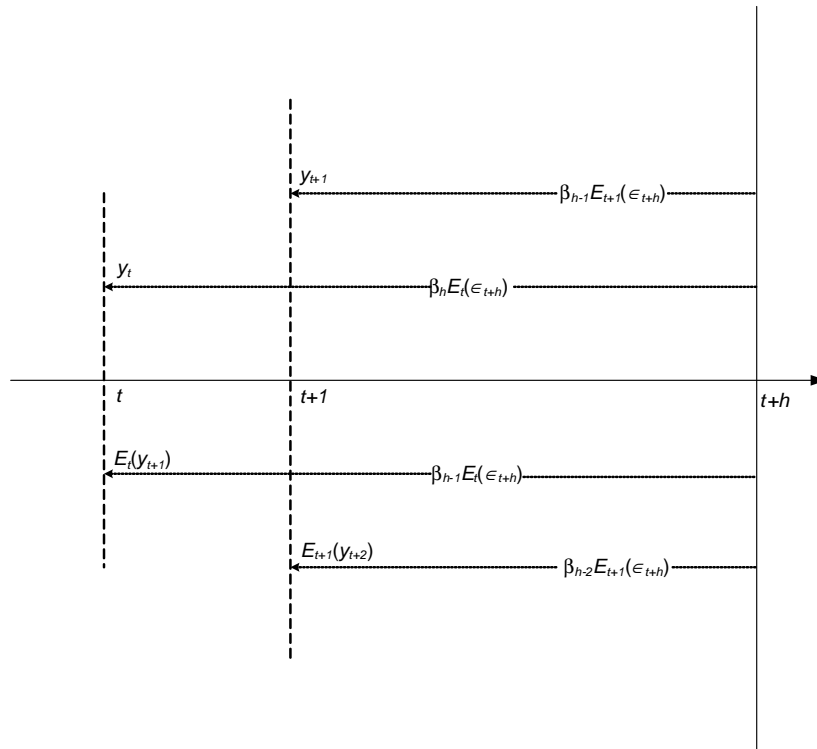


Figure 2: Rejection rates of the 5%-level LR test of  $H_0 : \phi_1 = 0.9$  for  $T = 100$  (solid line),  $T = 200$  (long dashes) and  $T = 500$  (dashes). The data are generated from a model with  $\phi_1 = 0.9 - c/\sqrt{T}$ .

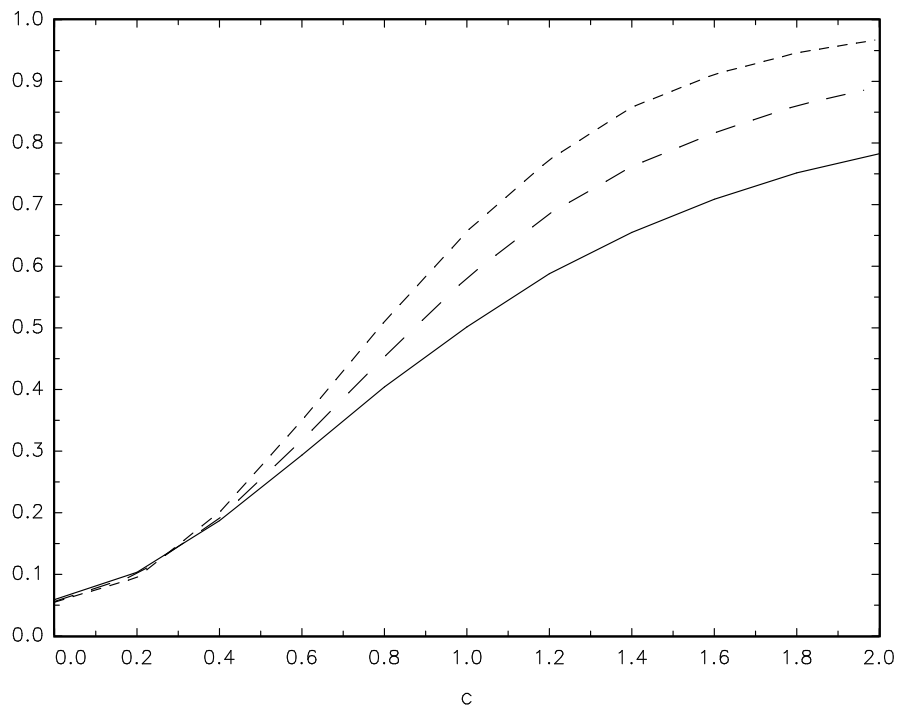


Figure 3: Autocorrelation function of the U.S. inflation. The dashed line depicts the upper bound of the approximate 95% confidence band.

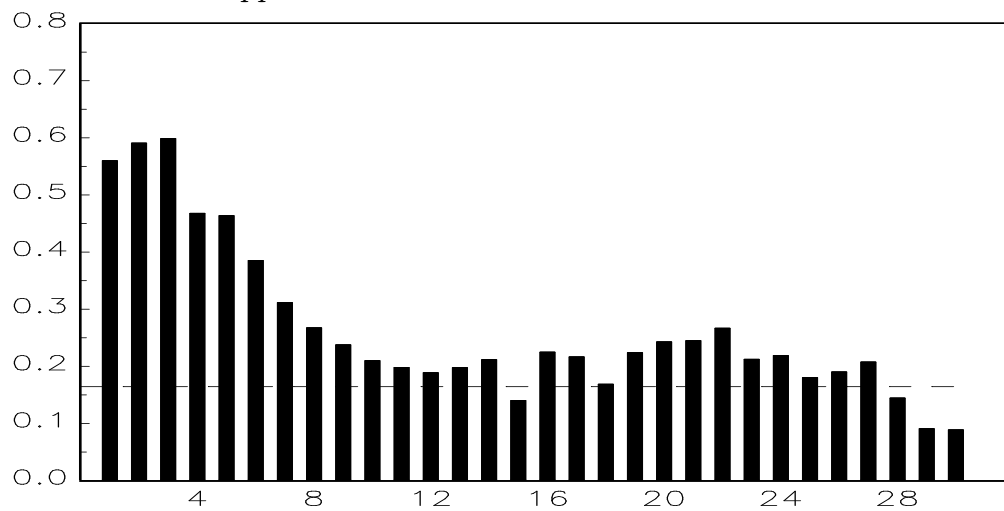


Figure 4: Quantile-quantile plots of the residuals of the AR(3,0)- $N$  and AR(0,3)- $t$  models.

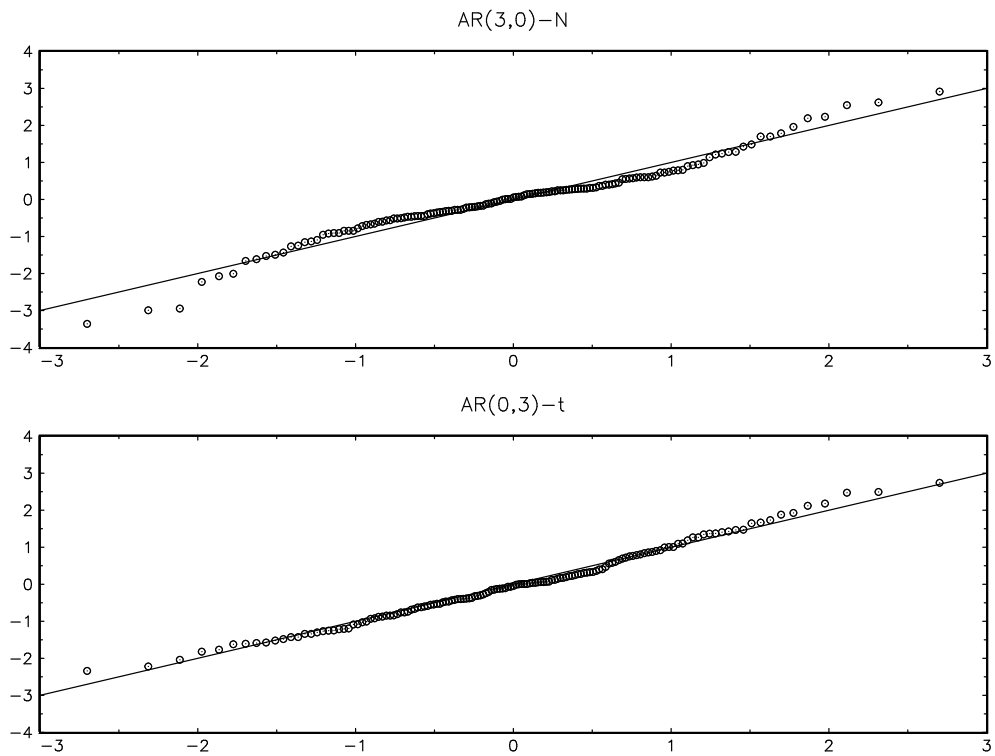


Table 1: Finite-sample properties of the ML estimator.

		DGP					
		$\phi_1=0.9, \varphi_1=0.9$		$\phi_1=0.9, \varphi_1=0.1$		$\phi_1=0.1, \varphi_1=0.9$	
$T$	Parameter	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
100	$\phi_1$	0.882	0.048	0.871	0.063	0.108	0.097
	$\varphi_1$	0.874	0.051	0.107	0.095	0.869	0.064
200	$\phi_1$	0.892	0.032	0.888	0.030	0.101	0.058
	$\varphi_1$	0.888	0.033	0.102	0.058	0.888	0.029
500	$\phi_1$	0.897	0.019	0.896	0.016	0.100	0.035
	$\varphi_1$	0.896	0.019	0.100	0.035	0.896	0.016

The DGP is the AR(1,1) model where the error term follows the  $t$ -distribution with 3 degrees of freedom and  $\sigma = 0.1$ . The results are based on 10,000 realizations.

Table 2: Rejection rates of the Wald and likelihood ratio (LR) tests.

		DGP					
		$\phi_1=0.9, \varphi_1=0.9$		$\phi_1=0.9, \varphi_1=0.1$		$\phi_1=0.1, \varphi_1=0.9$	
$T$	Parameter	Wald test	LR test	Wald test	LR test	Wald test	LR test
100	$\phi_1$	0.134	0.059	0.137	0.075	0.119	0.061
	$\varphi_1$	0.152	0.075	0.114	0.055	0.147	0.083
200	$\phi_1$	0.111	0.056	0.101	0.062	0.092	0.054
	$\varphi_1$	0.121	0.061	0.087	0.052	0.106	0.062
500	$\phi_1$	0.088	0.055	0.072	0.055	0.072	0.051
	$\varphi_1$	0.093	0.058	0.073	0.053	0.077	0.056

See notes to Table 1. The figures are rejection rates of Wald and LR tests of the null hypothesis that the parameter equals the true value. The nominal size of the tests is 5%.

Table 3: Simulation results on model selection by maximizing the likelihood function.

		DGP								
		$\phi_1=0.9, \varphi_1=0.9$			$\phi_1=0.9, \varphi_1=0.1$			$\phi_1=0.1, \varphi_1=0.9$		
$T$		AR(2,0)	AR(1,1)	AR(0,2)	AR(2,0)	AR(1,1)	AR(0,2)	AR(2,0)	AR(1,1)	AR(0,2)
100		765	8077	1158	844	5472	3684	3608	5402	990
200		205	9463	332	219	6806	2975	3034	6699	267
500		5	9991	4	4	8501	1495	1458	8538	4

See notes to Table 1. Each figure indicates the number of times the model in question maximizes the likelihood function out of 10,000 realizations.

Table 4: Estimation results of the autoregressive models for the demeaned U.S. inflation.

	Model				
	AR(3,0)- $N$	AR(3,0)- $t$	AR(2,1)- $t$	AR(1,2)- $t$	AR(0,3)- $t$
$\phi_1$	0.204 (0.080)	0.250 (0.058)	0.783 (0.066)	-0.027 (0.056)	
$\phi_2$	0.311 (0.078)	0.282 (0.058)	0.106 (0.066)		
$\phi_3$	0.311 (0.081)	0.303 (0.058)			
$\varphi_1$			-0.436 (0.060)	0.339 (0.050)	0.235 (0.050)
$\varphi_2$				0.448 (0.049)	0.383 (0.047)
$\varphi_3$					0.237 (0.050)
$\sigma$		1.987 (0.439)	2.060 (0.518)	2.168 (0.709)	2.114 (0.767)
$\lambda$		3.098 (0.983)	2.979 (0.953)	2.688 (0.726)	2.633 (0.723)
Log-likelihood	-287.686	-279.810	-281.601	-276.803	-269.815
Ljung-Box (4)	0.570	0.494	0.001	0.001	0.278
McLeod-Li (4)	0.004	0.004	0.057	0.281	0.873

AR( $r, s$ ) denotes the autoregressive model with the  $r$ th and  $s$ th order polynomials  $\phi(B)$  and  $\varphi(B^{-1})$ , respectively.  $N$  and  $t$  refer to Gaussian and  $t$ -distributed errors, respectively. The figures in parentheses are standard errors. Marginal significance levels of the Ljung-Box and McLeod-Li tests with 4 lags are reported.