Ensaio Econômico

EPGE

Escola

Brasileira de

Economia e

Finanças

Nº 810

ISSN 0104-8910

Mixed Causal-Noncausal Autoregressions with Exogenous Regressors

Alain Hecq, João Victor Issler, Sean Telg

Outubro de 2019

URL: <https://epge.fgv.br>

Os artigos publicados são de inteira responsabilidade de seus autores. As opiniões neles emitidas não exprimem, necessariamente, o ponto de vista da Fundação Getulio Vargas.

EPGE Escola Brasileira de Economia e Finanças

Diretor Geral: Rubens Penha Cysne

Vice-Diretor: Aloisio Araujo

Coordenador de Regulação Institucional: Luis Henrique Bertolino Braidó

Coordenadores de Graduação: Luis Henrique Bertolino Braidó & André Arruda Villela

Coordenadores de Pós-graduação Acadêmica: Humberto Moreira & Lucas Jóver Maestri

Coordenadores do Mestrado Profissional em Economia e Finanças: Ricardo de Oliveira Cavalcanti & Joísa Campanher Dutra

Hecq, Alain

Mixed Causal-Noncausal Autoregressions with Exogenous Regressors/ Alain Hecq, João Victor Issler, Sean Telg - Rio de Janeiro : FGV,EPGE, 2019

55p. - (Ensaio Econômico; 810)

Inclui bibliografia.

CDD-330

Mixed Causal-Noncausal Autoregressions with Exogenous Regressors

Alain Hecq* João Victor Issler† Sean Telg‡

October 9, 2019

Abstract

The mixed causal-noncausal autoregressive (MAR) model has been proposed to estimate time series processes involving explosive roots in the autoregressive part, as it allows for stationary forward and backward solutions. Possible exogenous variables are substituted into the error term to ensure the univariate MAR structure of the variable of interest. To study the impact of fundamental exogenous variables directly, we instead consider a MARX representation which allows for the inclusion of exogenous regressors. We argue that, contrary to MAR models, MARX models might be identified using second-order properties. The asymptotic distribution of the MARX parameters is derived assuming a class of non-Gaussian densities. We assume a Student's t -likelihood to derive closed form solutions of the corresponding standard errors. By means of Monte Carlo simulations, we evaluate the accuracy of MARX model selection based on information criteria. We examine the influence of the U.S. exchange rate and industrial production index on several commodity prices.

Keywords: mixed causal-noncausal process, non-Gaussian errors, identification, commodity prices. **JEL codes:** C22, E31, E37.

*Alain Hecq, Maastricht University, School of Business and Economics, Department of Quantitative Economics, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: a.hecq@maastrichtuniversity.nl.

†João Victor Issler, Graduate School of Economics - EPGE, Getulio Vargas Foundation, Praia de Botafogo 190s. 1100, Rio de Janeiro, RJ 22250-900, Brazil. E-mail: Joao.Issler@fgv.br.

‡Corresponding author: Sean Telg, Vrije Universiteit Amsterdam, Department of Econometrics, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. E-mail: j.m.a.telg@vu.nl. Tel: +31 20 59 89449.

This work was partly written while Sean Telg visited the CREST in Paris, Alain Hecq EPGE/FGV in Rio de Janeiro, and João Victor Issler Maastricht University. We thank all institutions for hosting us. We would like to express gratitude to Christian Francq and Jean-Michel Zakoïan for stimulating and fruitful discussions. We also thank participants of CFE (Sevilla, 2016), SNDE (Paris, 2017), EcoSta (Hong Kong, 2017), IAAE (Sapporo, 2017) and ESEM (Lisbon, 2017) as well as two anonymous referees for valuable comments and remarks. João Victor Issler thanks the financial support of CNPq, FAPERJ and CAPES on different grants. This study was partly financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) - Finance Code 001

1 Introduction

Beyond the standard argument that reversing time is helpful to obtain a stationary solution for explosive processes (Brockwell and Davis, 1991), the usefulness of mixed causal-noncausal autoregressive (MAR) models, i.e. time series specifications with both lags and leads components, can be explained by at least three additional reasons. First, Gouriéroux and Zakoïan (2016), Hencic and Gouriéroux (2014), Fries and Zakoïan (2019) and Hecq et al. (2016) demonstrate how the inclusion of noncausal autoregressive terms can generate dynamic patterns like speculative bubbles and asymmetric cycles that instead could only be generated using complex nonlinear models. Second, Lanne et al. (2012a, 2012b) show that allowing for noncausality might improve forecast performances. Third, the MAR representation of an economic variable can be seen as solutions of rational expectation models which are prevalent in macroeconomics (Lanne and Saikkonen, 2011). Whereas causal autoregressive models take only the fundamental solution of a system into account, the MAR model explicitly allows for nonfundamental outcomes, as the process of interest might depend on past, current and future (nonfundamental) shocks. This extension proves extremely useful, as it is well-known that some economic models do not possess a fundamental solution by construction (Alessi et al., 2011) and explicitly consider that the information available to economic agents is larger than that possessed by econometricians.

The inclusion of exogenous regressors proposed in this paper further improves the relevance of the MAR specification for modelling and forecasting economic processes. This model, enriched with exogenous variables, is denoted MARX. Our motivation for introducing covariates in MAR regressions is threefold. First, the MARX model offers more flexibility, as the addition of exogenous components with potentially marginal leptokurtic distributions, enriches the nonlinear patterns. The model can be seen as a generalization of not only the autoregressive model with exogenous regressors (ARX), but also the class of autoregressive distributed lag (ARDL) models and marginal equations of a class of VAR models. In terms of interpretation,

practitioners are able to study directly the impact of exogenous forcing variables on the variable of interest. In the existing literature (see e.g. Lanne and Luoto, 2013), exogenous regressors are included in the error term on which an MAR structure is assumed. Whereas this might be beneficial in terms of modelling structural equations, it hides the direct link between key economic variables. Second, the MARX model allows us to investigate the claim that when controlling for additional variables, the lead part might become insignificant. The reason for this is that it has been suggested that the noncausal component acts as a proxy for the omission of important variables that are not included in the model (Lof, 2013; Lof and Nyberg, 2017). Since both noncausal and causal models with exogenous regressors are contained in the class of MARX models, we can study conditions under which the practitioner favours the one specification over the other. Third, MARX processes have the interesting feature that, contrary to MAR processes, one can discriminate between causal, noncausal and mixed specifications based on second-order properties. The introduction of cross-covariances leads to asymmetries in the autocovariance functions of the different models. This allows practitioners to perform a confirmatory analysis. In the empirical application, we consider commodity prices that are expected *(i)* to show signs of noncausality (see e.g. Lof and Nyberg, 2017; Karapanagiotidis, 2014) due to their forward-looking nature and *(ii)* are likely to depend on an indicator of economic activity (Issler et al., 2014) and the U.S. dollar exchange rate.

The remainder of this paper is organized as follows. In Section 2, we introduce the MARX model and illustrate how to simulate such processes. We propose a model selection and forecasting procedure and show that MARX models have the appealing feature to be (potentially) identifiable by second-order properties. Section 3 shows how MARX parameters can be estimated by maximum likelihood (ML) for non-Gaussian densities as defined in Andrews et al. (2006). For the case of the Student's t -likelihood, we provide a method to compute closed form solutions of the corresponding standard errors. By means of Monte Carlo simulations, we evaluate in Section 4 the performance of the ML estimator, our proposed model selection method, the identifiability

of MARX processes based on second-order properties and the forecasting performance of MARX models. Section 5 consists of the empirical application in which we highlight the main features of the MARX model and their usefulness for practitioners. Section 6 summarizes and concludes. The Appendix contains details on the properties of the approximate ML estimator. Proofs and additional material are collected in the Supplementary Material. The methods proposed in this paper are implemented in the R package **MARX**; its specifications are discussed in detail in Hecq et al. (2017a). The package is freely available at <https://CRAN.R-project.org/package=MARX>.

2 The MARX Model

Let y_t be the variable of interest which is observed over the time period $t = 1, \dots, T$. Let $x_{i,t}$ ($i = 1, \dots, q$) be the i th variable in a set of q explanatory regressors and $\beta \in \mathbb{R}^q$ a vector of parameters. Then we can define $\mathbf{X}_t = [x_t, \dots, x_{q,t}]' \in \mathbb{R}^q$ as the vector of all exogenous variables at time t .¹ The MARX(r, s, q) for a stationary time series y_t can be represented as

$$\phi(L)\varphi(L^{-1})y_t - \beta'\mathbf{X}_t = \varepsilon_t, \quad (1)$$

where $\phi(L)$ is a lag polynomial of order r , $\varphi(L^{-1})$ a lead polynomial of order s and we denote $r + s = p$. The operator L is the lag operator when raised to positive powers, i.e., $L^i y_t = y_{t-i}$, and interpreted as a lead operator when raised to negative powers: $L^{-i} y_t = y_{t+i}$. All roots of $\phi(z)$ and $\varphi(z^{-1})$ lie outside the unit circle.² The sequence ε_t is assumed to be *iid* non-Gaussian. In order to ensure identifiability of the parameter vector β and to prove consistency of the maximum likelihood (ML) estimator, we make the following assumptions on $\mathbf{X}_t$:

Assumptions. The processes in $\mathbf{X}_t$ are

(I) ergodic, (strictly) stationary and strictly exogenous w.r.t. ε_t , i.e. $\mathbb{E}(\varepsilon_t | x_{i,t-j}) = 0$ for all

¹ For now, we only consider contemporaneous values of $\mathbf{X}_t$ in the model. However, the MARX model can also take the form of a mixed autoregressive distributed lag (MARDL) model.

² We replace the operator L by the complex variable z when considering the properties of polynomials.

$i = 1, \dots, q$ and all $j \in \mathbb{Z}$;

(II) of the form $x_{i,t} = c_i + \sum_{j=-\infty}^{\infty} \rho_{i,j} \eta_{i,t-j}$ with $\eta_{i,t} \stackrel{iid}{\sim} (0, \sigma_{\eta_i}^2)$ for all $i = 1, \dots, q$;

(III) linearly independent.

Assumption (I) is necessary for the Central Limit Theorem for m -dependent processes (Theorem 6.4.2 in Brockwell and Davis, 1991). Whereas this zero conditional mean assumption restricts the economic relationships to be studied, we consider it as a natural extension to the exogeneity assumption imposed on the well-known ARX model. Assumption (II) defines the dynamic structure the processes in $\mathbf{X}_t$ can assume. By allowing for a two-sided moving average process, purely causal and noncausal as well as mixed ARMA processes are included. By assumption (III) the matrix of explanatory variables has full column rank.

Note that the concepts of causality and noncausality are defined in terms of the strictly stationary solution of the model. When $q = 0$, the process in (1) reduces to an MAR process that has a two-sided moving average representation consisting of past, current and future values of ε_t under appropriate moment conditions for ε_t and summability conditions for the sequence of moving average coefficients (see, e.g. Gouriéroux and Zakoïan, 2016, for the infinite variance case). In case $q > 0$, the process considered no longer has a strictly stationary solution solely in terms of ε_t , but involves both $\mathbf{X}_t$ and ε_t . That is,

$$y_t = \pi(L, L^{-1})[\varepsilon_t + \beta' \mathbf{X}_t] = \sum_{j=-\infty}^{\infty} \pi_j z_{t-j}, \quad (2)$$

where $z_{t-j} = \varepsilon_{t-j} + \sum_{i=1}^q \beta_i x_{i,t-j}$ and $\pi(z, z^{-1})$ is a polynomial satisfying $\pi(z, z^{-1})\phi(z)\varphi(z^{-1}) = 1$. The process y_t exists almost surely under rather weak conditions, such as $\sum_{j=-\infty}^{\infty} |\pi_j|^\delta < \infty$ and $\mathbb{E}|z_t|^\delta < \infty$ for $\delta > 0$, where the polynomials $\phi(z)$ and $\varphi(z^{-1})$ are invertible (Gouriéroux and Jasiak, 2016). We observe that y_t still has a two-sided moving average representation, but augmented with a second part involving linear combinations of past, current and future values of $\mathbf{X}_t$. Since $\beta_i x_{i,t}$ can be interpreted as a new series $\tilde{x}_{i,t}$ which is the series $x_{i,t}$ multiplied

by a constant term β_i , y_t in (2) consists of two additive parts: (i) a two-sided moving average representation and (ii) the sum of q processes $\tilde{x}_{i,t}$ that are passed through a two-sided linear filter with coefficients resulting from inverting the product $[\phi(z)\varphi(z^{-1})]$ to $\pi(z, z^{-1})$.³

In order to make a comparison of the MARX and MAR model, we have to be more explicit on the role of $\mathbf{X}_t$ in (1). According to Assumption (II), the exogenous regressors have a representation as a two-sided moving average. Hence, let us assume z_t can be written as a mixed causal-noncausal process, i.e.

$$\rho(L)\gamma(L^{-1})z_t = \varepsilon_t^*, \quad (3)$$

where ε_t^* is an *iid* error term. After inverting the polynomials, we can substitute z_t for $\beta' \mathbf{X}_t + \varepsilon_t$ in (1) to obtain

$$\phi(L)\varphi(L^{-1})y_t = [\rho(L)\gamma(L^{-1})]^{-1}\varepsilon_t^*,$$

which yields

$$\phi^*(L)\varphi^*(L^{-1})y_t = \varepsilon_t^*,$$

with $\phi^*(L) \equiv \phi(L)\rho(L)$ and $\varphi^*(L^{-1}) \equiv \varphi(L^{-1})\gamma(L^{-1})$. Hence, if we want to represent the MARX by an MAR model, the dynamics of z_t , which are determined by the fluctuations in $\mathbf{X}_t$, might increase the degree of the causal and noncausal polynomial of the MAR. In other words, given the dynamics of z_t , we have that $\deg[\phi^*(z)] \geq \deg[\phi(z)]$ and $\deg[\varphi^*(z)] \geq \deg[\varphi(z)]$.

Lemma 1. *From (1), we can construct the unobserved noncausal and causal components (u, v) similar to Lanne and Saikkonen (2011) and Gouriéroux and Jasiak (2016) and obtain:*

$$u_t \equiv \phi(L)y_t \leftrightarrow \varphi(L^{-1})u_t - \beta' \mathbf{X}_t = \varepsilon_t, \quad (4)$$

$$v_t \equiv \varphi(L^{-1})y_t \leftrightarrow \phi(L)v_t - \beta' \mathbf{X}_t = \varepsilon_t. \quad (5)$$

³The effects of two-sided linear filters, with a focus on seasonal adjustment, on the identification of mixed causal-noncausal models is studied in Hecq et al. (2017b).

Defining $[\varphi(z^{-1})]^{-1} \equiv \delta(z^{-1})$ and $[\phi(z)]^{-1} \equiv \alpha(z)$, the second equalities in (4)-(5) can be written as $u_t = \sum_{j=0}^{\infty} \delta_j z_{t+j}$ and $v_t = \sum_{j=0}^{\infty} \alpha_j z_{t-j}$, which explains why u_t and v_t are denoted unobserved noncausal and causal components respectively. The corresponding first equations show that the filtered values (u, v) can be interpreted as a noncausal (resp. causal) “error term” of a purely causal (resp. noncausal) autoregression.

2.1 Simulation of MARX Processes

The filtered values defined in Lemma 1 establish a deterministic dynamic relationship between the unobserved components u_t and v_t , the exogenous variables $\mathbf{X}_t$ and the process y_t , which can be used to simulate various MARX(r, s, q) series. Similar to Gouriéroux and Jasiak (2016), we make use of the independence of specific blocks of u, v and y values. The main difficulty for generating MARX(r, s, q) is the product of polynomials $\phi(z)\varphi(z^{-1})$. One cannot directly simulate such a process as simultaneously initial and terminal values are required. If the degree of (at least) one of the polynomials equals 0, the problem is greatly simplified.⁴ In the general MARX(r, s, q) setup, filtered values are used to circumvent the problem.

We characterize the simulation procedure as a two-step approach. First, note that u_t can be simulated directly using the second equality of (4) by taking s terminal values, say $u_{T+1}^*, \dots, u_{T+s}^*$ and simulating sequences for ε_t and $\mathbf{X}_t$. Second, y_t can be constructed using the first equality of (4) similar to a conventional causal autoregressive process. To simulate this process, one needs r starting values for y_t , say $y_{-1}^*, \dots, y_{-r}^*$. It is highly advised to allow for a burn-in period in both steps to remove the dependence on the terminal values of u_t and the starting values of y_t . Note that (5) could equivalently be used to simulate an MARX process.

[FIGURE 1 AROUND HERE]

Figure 1 shows simulated paths of an MARX(1,1,1) and MAR(1,1) process with $[\phi_1, \varphi_1]' =$

⁴In that case y_t can easily be simulated directly by generating a sequence of ε_t and choosing starting [terminal] values for y_t and $\mathbf{X}_t$ in the causal [noncausal] case.

$[0.3, 0.9]'$, $\beta_1 = 0.3$, $x_t \stackrel{iid}{\sim} t(1, 1)$ and $\varepsilon_t \stackrel{iid}{\sim} t(3, 1)$ with $t(\nu, \sigma)$ being the Student's t distribution with degrees of freedom parameter ν and scale parameter σ . It can be seen that both processes generally move similarly with the major exception that the MARX process contains more peaks and troughs, which are also more extreme in comparison. This is due to the choice of x_t , which is, solely for expository purposes, chosen to be standard Cauchy distributed. Hence, the MARX specification takes into account shocks that cannot be explained by past and future values of the dependent variable, but which are present because of major changes in explanatory variables at some specific points in time.

2.2 Interpretation of the MARX Model

The MARX model allows practitioners to study the direct impact of exogenous regressors on the variable of interest. We argue that there are more reasons to study the MARX model. In particular, consider the autoregressive distributed lag (ARDL) model of the form⁵

$$y_t = \sum_{i=1}^p a_i y_{t-i} + \sum_{i=0}^k \mathbf{b}_i' \mathbf{X}_{t-i} + \epsilon_t. \quad (6)$$

We now show that the MARX model allows for different representations of (6) which can exhibit richer dynamics. Define $a(L) \equiv (1 - \sum_{i=1}^p a_i L^i)$ and suppose, similarly to Breidt et al. (1991), that the p th order polynomial $a(z)$ has r well-behaved roots (i.e., outside the unit circle) and s ill-behaved roots (i.e. inside the unit circle) with $p = r + s$. Following Lanne and Saikkonen (2011), write $a(L) = \phi(L)\varphi^*(L)$ with $\phi(z)$ and $\varphi^*(z)$ being polynomials of orders r and s respectively. Now we note that the polynomial $\phi(z)$ has the well-behaved roots, whereas $\varphi^*(z)$ possesses explosive roots in direct time. Considering $\varphi^*(z)$ in reverse time, all roots are inverted and thus outside the unit circle. Straightforward algebra gives:

$$\varphi^*(z) = -\varphi_s^* z^s \varphi(z^{-1}), \quad (7)$$

⁵For notational convenience, we fix the lag order k to be the same for all variables contained in $\mathbf{X}_t$.

which yields,

$$\phi(L)\varphi(L^{-1})y_t = \left[-\frac{1}{\varphi_s^*} L^{-s} \right] \left(\sum_{i=0}^k \mathbf{b}'_i \mathbf{X}_{t-i} + \epsilon_t \right).$$

Denoting $\varepsilon_t = \left(-\frac{1}{\varphi_s^*} \epsilon_{t+s} \right)$ and $\beta'_i = \left(-\frac{1}{\varphi_s^*} \mathbf{b}'_i \right)$, we obtain

$$\phi(L)\varphi(L^{-1})y_t = \sum_{i=0}^k \beta'_i \mathbf{X}_{t-i+s} + \varepsilon_t, \quad (8)$$

which is a form of (1) explicitly allowing for the inclusion of lags and leads of the exogenous regressors. As explained in footnote 3, the MARX model is allowed to include such terms and can be labeled as MARDL. Lanne and Saikkonen (2011) discuss the condition ($\varphi_s \neq 0$) for which a one-to-one correspondence between (6) and (8) exists for the MAR model. The latter specification has the same advantages in the MARX case: it is easier to set up conventional likelihood tests to determine r and s and the model parameters are orthogonal to the distributional parameters, which implies asymptotic independence of the corresponding ML estimates.

Example 1. Consider the process $y_t = a_1 y_{t-1} + a_2 y_{t-2} + b_1 x_{t-1} + \epsilon_t$, which is nested in (6). Using equation (7), following the algebra above, this model can equivalently be written as a MARX model of the form

$$(1 - \phi_1 L)(1 - \varphi_1 L^{-1})y_t = \beta_1 x_t + \varepsilon_t. \quad (9)$$

The MARX (or more specifically, MARDL) allows one to consider the noncausal and mixed causal-noncausal counterparts of the set of models in (6). Nested within this specification are models that consider Granger-causality from x_t to y_t (as in Example 1) and single equations within a VAR. That is, if we extend the process in Example 1 with the term $b_2 x_{t-2}$, it can be seen as a marginal equation of a bivariate VAR(2) with uncorrelated error terms.

Example 2. *The VAR(2) model is given by*

$$\begin{aligned} y_t &= a_{11}y_{t-1} + a_{12}x_{t-1} + b_{11}y_{t-2} + b_{12}x_{t-2} + \epsilon_{y,t} \\ x_t &= a_{21}y_{t-1} + a_{22}x_{t-1} + b_{21}y_{t-2} + b_{22}x_{t-2} + \epsilon_{x,t} \end{aligned}$$

Note that the exogeneity assumption (I) implies that x_t is not determined by y_t at any displacement in time. Hence, the MARX model is compatible with a situation in which $a_{21} = b_{21} = 0$ and $\text{Cor}(\epsilon_{y,t}, \epsilon_{x,t}) = 0$. Under these restrictions, the MARX represents the first equation of the VAR. Whenever the exogeneity assumption cannot be credibly made, a better modelling approach might be to compare causal, mixed causal-noncausal and purely noncausal VAR models.

Another observation can be made from Example 1. In order to compare causal, noncausal and mixed models that are theoretically equivalent, the time subscript of the exogenous regressor has to be adjusted by the degree of noncausality. That is, the ARDL considers x_{t-1} , whereas the MARX model contains x_t as the noncausal order is increased by one. This consideration might be important in cases of structural modelling, as it is crucial to identify the corresponding “correct” reduced form. In terms of prediction, the main point of interest is to select a model that fits the data best. In Section 5, we observe that no time shift of the exogenous regressors typically results in higher log-likelihood values.

Lastly, we show that the MARX is compatible with the rational-expectation models studied in Lanne and Luoto (2013), with a caveat. After some algebra, (9) can be rewritten as

$$y_t = \frac{\phi_1}{1 + \phi_1\varphi_1}y_{t-1} + \frac{\varphi_1}{1 + \phi_1\varphi_1}y_{t+1} + \frac{\beta_1}{1 + \phi_1\varphi_1}x_t + \varepsilon_t. \quad (10)$$

Calling these coefficients γ_f, γ_b and λ respectively and by replacing y_{t+1} by $E(y_{t+1})$ plus a martingale difference sequence error term, we obtain a specification similar to Lanne and Luoto (2013, Equation 4) expressed in regression form.

The caveat alluded above is that a MARX model must obey the restriction imposed by Assumption (I), i.e., variable x_t must be strictly exogenous. Thus, unless we already start with a strictly exogenous x_t in our analysis, a MARX representation is not obtained in the end. It is important to stress that there are plausible examples covered by the rational-expectation literature where Assumption (I) may fail. But, this is not true in general, and whether one can obtain a MARX model at the end will be essentially an empirical question, guided by economic theory. In Section 5 we present empirical applications where Assumption (I) is credibly matched.

2.3 Model Selection

The findings in the previous section provide important considerations to perform model selection in the framework of MARX models. We present the model selection procedure proposed in this paper and illustrate it by an example afterwards.

- **Step 1: Determining total autoregressive order.** In a first step, we identify the so-called pseudo-causal model (see e.g. Hecq et al., 2016) by estimating ARDL models as in (6) by OLS and identify by information criteria like AIC, BIC and HQ the autoregressive order p that makes the residuals free of serial correlation.⁶ Additionally, we find the lowest lag orders of the exogenous regressors to consider in the estimation of all models. No leads can be selected in the ARDL, as it violates the exogeneity condition in Assumption (I).
- **Step 2: From causal to noncausal.** From the identified ARDL model, we estimate the theoretically equivalent mixed and noncausal models with $p = r + s$ where the time subscript of the exogenous regressors is adjusted by the degree of noncausality. The proposed estimation method is the non-Gaussian approximate maximum likelihood (AML) estimator which is discussed in Section 3. As the exogeneity condition is respected in Step

⁶It is advised to perform diagnostic tests to see whether additional lags are needed to remove autocorrelation from the series. The null of normality should be tested on the residuals of the pseudo-causal model to justify the non-Gaussian ML estimator and to look for signs of noncausality. A description of the model selection procedure for MAR models can be found in Hecq et al. (2016).

1, it also holds here and controls the maximum allowed time shift in the noncausal model.

- **Step 3: From noncausal to causal.** We estimate a purely noncausal model of order p with contemporaneous exogenous regressors together with all theoretically equivalent MARX models with $p = r + s$ by non-Gaussian AML.
- **Step 4: Remaining models.** The models in Step 2 and 3 are MARX models that consider the exogenous regressors at different points in time. At this step, we propose to estimate models which consider the exogenous regressors at time periods in between the ones found at Step 2 and 3 by non-Gaussian AML.
- **Step 5: Highest log-likelihood.** The model that attains the highest value for the log-likelihood at its estimated parameters is chosen to be the final model. The set of considered models are the ones estimated in Step 2, 3 and 4 as well as the original ARDL which is re-estimated by AML.

Compared to the MAR model, the MARX model selection procedure is more complicated. The first step for the MAR consists of estimating causal AR models by OLS, which is solely important to consistently estimate the lag order that makes the residuals free of serial correlation. In the MARX case, the significance of the exogenous regressors might change depending on the type of model. For example, the regressors might be significant in causal models, while they are not in the noncausal model. To take such situations into account, Step 2 and 3 consider the causal and noncausal model as separate starting points. Step 4 examines the remaining models for the sake of completeness. Thus, these extra steps take into account that exogenous regressors create asymmetries based on the direction of timeline considered. The procedure is easily characterized by an example.

Example 3. *Consider the case of one exogenous regressor x_t . Suppose we identify the following ARDL model using information criteria in the first step: y_t on y_{t-1}, y_{t-2}, x_t . Table 1 displays the proposed model selection.*

Step 1	Step 2	Step 3
Selected ARDL model: C: y_t on y_{t-1}, y_{t-2}, x_t	Estimate by non-Gaussian AML theoretically equivalent models: M: y_t on $y_{t-1}, y_{t+1}, x_{t+1}$ NC: y_t on $y_{t+1}, y_{t+2}, x_{t+2}$	Noncausal starting point: NC: y_t on y_{t+1}, y_{t+2}, x_t Theoretically equivalent models: M: y_t on $y_{t-1}, y_{t+1}, x_{t-1}$ C: y_t on $y_{t-1}, y_{t-2}, x_{t-2}$
Step 4	Step 5	
Remaining models: C: y_t on $y_{t-1}, y_{t-2}, x_{t-1}$ M: y_t on y_{t-1}, y_{t+1}, x_t NC: y_t on $y_{t+1}, y_{t+2}, x_{t+1}$	Re-estimate ARDL of Step 1 by non-Gaussian AML and compare log-likelihood of models in all steps.	→ Select model with highest log-likelihood
C, M and NC stand for causal, mixed and noncausal respectively		

Table 1: Model selection example

Remark 1. By extensive simulations, we find that if the coefficients of the exogenous regressors are zero in the true data generating process (DGP), then the OLS estimates of these coefficients are found to be insignificant. Consequently, we justify the estimation of the ARDL by OLS in the first step using results in Gouriéroux and Jasiak (2018). They show that OLS consistently estimates the lag order p , while the non-Gaussian ML estimator might underestimate the autoregressive order in the presence of heavy-tailed distributions such as the Cauchy.

2.4 Identifiability

Whereas we propose a non-Gaussian AML estimator to estimate MARX models, we argue that it is theoretically possible to distinguish between different specifications based on second-order properties. Contrary to the MAR model, the presence of cross-covariances in the MARX models allows to achieve identification without distributional assumptions. We characterize how this works in practice by considering first-order autoregressive processes first.

Example 4. Suppose the true DGP is the $MARX(1,0,1)$ given by $y_t = \phi_1 y_{t-1} + \beta_C x_t + \varepsilon_t$. The corresponding noncausal model to consider is the $MARX(0,1,1)$ which can be represented

as $y_t = \varphi_1 y_{t+1} + \beta_{NC} x_{t+1} + \epsilon_t$. In case of the MARX(1,0,1) model, the OLS estimator fulfills

$$\begin{aligned}\hat{\phi}_{1,T} &= \frac{\sum_{t=r+1}^{T-s} y_t y_{t-1}}{\sum_{t=r+1}^{T-s} y_{t-1}^2} - \hat{\beta}_{C,T} \frac{\sum_{t=r+1}^{T-s} y_{t-1} x_t}{\sum_{t=r+1}^{T-s} y_{t-1}^2} \xrightarrow{p} \frac{\gamma_y(1)}{\gamma_y(0)} - \beta_C \frac{\gamma_{xy}(1)}{\gamma_y(0)}, \\ \hat{\beta}_{C,T} &= \frac{\sum_{t=r+1}^{T-s} x_t y_t}{\sum_{t=r+1}^{T-s} x_t^2} - \hat{\phi}_{1,T} \frac{\sum_{t=r+1}^{T-s} y_{t-1} x_t}{\sum_{t=r+1}^{T-s} x_t^2} \xrightarrow{p} \frac{\gamma_{xy}(0)}{\gamma_x(0)} - \phi_1 \frac{\gamma_{xy}(1)}{\gamma_x(0)},\end{aligned}$$

where $\gamma_x(k)$ and $\gamma_y(k)$ denote the autocovariance function of x_t and y_t at order k respectively. Additionally, $\gamma_{xy}(k)$ denotes the cross-covariance between x_t and y_{t-k} . For the MARX(0,1,1) model, results are only slightly different:⁷

$$\begin{aligned}\hat{\phi}_{1,T} &= \frac{\sum_{t=r+1}^{T-s} y_t y_{t+1}}{\sum_{t=r+1}^{T-s} y_{t+1}^2} - \hat{\beta}_{NC,T} \frac{\sum_{t=r+1}^{T-s} y_{t+1} x_{t+1}}{\sum_{t=r+1}^{T-s} y_{t+1}^2} \xrightarrow{p} \frac{\gamma_y(1)}{\gamma_y(0)} - \beta_{NC} \frac{\gamma_{xy}(0)}{\gamma_y(0)}, \\ \hat{\beta}_{NC,T} &= \frac{\sum_{t=r+1}^{T-s} x_{t+1} y_t}{\sum_{t=r+1}^{T-s} x_{t+1}^2} - \hat{\phi}_{1,T} \frac{\sum_{t=r+1}^{T-s} y_{t+1} x_{t+1}}{\sum_{t=r+1}^{T-s} x_{t+1}^2} \xrightarrow{p} \frac{\gamma_{xy}(1)}{\gamma_x(0)} - \varphi_1 \frac{\gamma_{xy}(0)}{\gamma_x(0)}.\end{aligned}$$

For the estimation of the autoregressive parameter, information about the cross-covariance between x_t and y_t is contained that is different for both specifications. Note that this is not due to the fact that x_t enters at different time periods in both models considered. If both models consider the exogenous variable at the same point in time, the OLS estimates still differ from one another as for cross-covariances $\gamma_{xy}(k) \neq \gamma_{xy}(-k)$ generally ($k = 1, 2, 3, \dots$). In the same spirit, Cubadda et al. (2019) show that reduced rank restrictions help to identify purely causal from purely noncausal VAR models in a Gaussian framework whereas unrestricted models are not identifiable by Gaussian MLE.

Discriminating between different specifications is likely to depend on sample size T and the importance of the exogenous regressor, i.e. the value of its coefficient. OLS estimation can be used whenever we restrict our attention to pure models. If the model has both a causal and noncausal part, OLS cannot be applied due to the multiplicative structure of the polynomials.

⁷In case of the MARX(0,1,1) without shifted the exogenous regressor, x_{t+1} should be replaced by x_t and $\gamma_{xy}(0)$ by $\gamma_{xy}(-1)$ in both equations below.

In that case, we propose the Gaussian ML estimator.

Remark 2. A problem of model identification by second-order properties might occur when the set of exogenous regressors is not significant in all time directions. For example, suppose a purely noncausal model (without exogenous regressors) is the true DGP. It can be shown that the residual sum of squares of purely causal and noncausal models without exogenous regressors are asymptotically equal. Since noncausal models might create time series with heteroskedasticity in direct time (Gouriéroux and Zakoïan, 2016), the inclusion of volatile exogenous regressors in causal models might improve the fit of the model. Based on the measure sum of squared residuals (RSS), the misspecified causal model with covariates might be preferred. However, as the number of regressors differ between causal and noncausal models in such a situation, it seems better to consider a measure that takes this into account, e.g. information criteria.

The discussion can be generalized to processes with higher order dynamics and multiple regressors with possibly infinite unconditional mean and/or variance.⁸ In this paper, assumption (II) excludes those processes from the analysis. Since the identifiability of the MARX model by either OLS or Gaussian MLE depends on the significance of the coefficients of the exogenous regressors, we propose to (initially) use a non-Gaussian ML estimator.

3 Estimation and Forecasting

3.1 Approximate Maximum Likelihood Estimation

Maximum likelihood estimation of noncausal autoregressive models has been studied by Breidt et al. (1991), Andrews et al. (2006) and Lanne and Saikkonen (2011).⁹ They show that ML

⁸Gouriéroux and Zakoïan (2016) show that sample autocorrelations are consistent for noncausal processes with α -stable error term, despite the fact that theoretical autocorrelations do not exist.

⁹Breidt et al. (1991) specify a noncausal model as a conventional autoregressive model that has roots inside the unit circle, while Andrews et al. (2006) consider all-pass models which are widely used in fitting noncausal autoregressions. Lanne and Saikkonen (2011) have a similar model setup to (1), the only difference being the exclusion of exogenous variables $\mathbf{X}_t$.

estimators are consistent and asymptotically normal under general conditions. This section builds on these results and establishes similar results for mixed causal-noncausal autoregressive models with exogenous regressors.

Similar to Lanne and Saikkonen (2011), we assume that the density function $f_\sigma(x; \boldsymbol{\lambda}) = \sigma^{-1}f(\sigma^{-1}x; \boldsymbol{\lambda})$ satisfies the regularity conditions of Andrews et al. (2006).¹⁰ The permissible parameter space of $\boldsymbol{\lambda}$, which collects the distributional parameters (except for the scale parameter), is denoted by $\boldsymbol{\Lambda}$, and is some subset of $\mathbb{R}^d$. The scale parameter σ only takes positive values, i.e., $\sigma > 0$. The permissible space of the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$ is defined by the stationarity condition, i.e. $\phi(z)$ and $\varphi(z)$ have all roots outside the unit circle. Using the independence of the blocks $(v_1, \dots, v_r)$, $(\varepsilon_{r+1}, \dots, \varepsilon_{T-s})$ and $(u_{T-s+1}, \dots, u_T)$, the density of the process y_t can be written as the product of the densities of these three variables (see Supplementary Material for details). However, since (the densities of) the first and third block do not depend on sample size T , we can approximate the density of y_t by the density of the second block. Replacing ε_t by the left-hand side of (1) and taking logs, we obtain the following log-likelihood function:

$$L_T(\boldsymbol{\theta}) = \sum_{t=r+1}^{T-s} \ln f_\sigma(\phi(L)\varphi(L^{-1})y_t - \boldsymbol{\beta}'\mathbf{X}_t; \boldsymbol{\lambda}) = \sum_{t=r+1}^{T-s} g_t(\boldsymbol{\theta}), \quad (11)$$

where $\boldsymbol{\theta} = [\boldsymbol{\phi}', \boldsymbol{\varphi}', \boldsymbol{\beta}', \boldsymbol{\lambda}', \sigma]'$. For convenience, we denote the ‘approximate’ sample size used to compute the log-likelihood $(T - p)$ by n . We can use the definition of the filtered values in (4) and (5) to write the series u_t and v_t as functions of the parameters, i.e., $u_t(\boldsymbol{\phi})$ and $v_t(\boldsymbol{\varphi})$. Then we can characterize $g_t(\boldsymbol{\theta})$ as follows:

$$\begin{aligned} g_t(\boldsymbol{\theta}) &= \ln f(\sigma^{-1}(v_t(\boldsymbol{\varphi}) - \phi_1 v_{t-1}(\boldsymbol{\varphi}) - \dots - \phi_r v_{t-r}(\boldsymbol{\varphi}) - \boldsymbol{\beta}'\mathbf{X}_t); \boldsymbol{\lambda}) - \ln(\sigma) \\ &= \ln f(\sigma^{-1}(u_t(\boldsymbol{\phi}) - \varphi_1 u_{t+1}(\boldsymbol{\phi}) - \dots - \varphi_s u_{t+s}(\boldsymbol{\phi}) - \boldsymbol{\beta}'\mathbf{X}_t); \boldsymbol{\lambda}) - \ln(\sigma). \end{aligned}$$

¹⁰The regularity conditions of Andrews et al. (2006) will henceforth be assumed. Densities that satisfy these conditions include a rescaled t -density and a weighted average of Gaussian densities.

Maximizing $L_T(\boldsymbol{\theta})$ over permissible values of $\boldsymbol{\theta}$ gives an approximate ML estimator of $\boldsymbol{\theta}$. We assume for now that the orders r and s are known. Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$ (and similarly for its components). Furthermore, assume that $\boldsymbol{\lambda}_0$, the true value of $\boldsymbol{\lambda}$, is an interior point of $\boldsymbol{\Lambda}$. Similar to Lanne and Saikkonen (2011), we can derive the asymptotic distribution of the approximate maximum likelihood estimator. Details are provided in Appendix A.

Lemma 2. *If conditions (A1)-(A7) of Andrews et al. (2006) and assumptions (I)-(III) hold, then*

$$\frac{1}{\sqrt{n}} \sum_{t=r+1}^{T-s} \frac{\partial}{\partial \boldsymbol{\theta}} g_t(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \text{diag}(\boldsymbol{\Sigma}, \boldsymbol{\Omega})).$$

Moreover, the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$, as defined in Appendix A, are positive definite.

The block diagonality of the covariance matrix of the limiting distribution follows directly from the formulation of the model in terms of both lag and lead operator. Hence, this result follows directly from Lanne and Saikkonen (2011) and Andrews et al. (2006). This ensures that both the autoregressive and exogenous (variables) parameters are orthogonal to the distributional parameters. The positive definiteness of $\boldsymbol{\Omega}$ is assumed through condition (A6) of Andrews et al. (2006). The positive definiteness of $\boldsymbol{\Sigma}$ follows, similar to the MAR case, from the condition $\mathcal{J} \equiv \int \frac{(f'(x; \boldsymbol{\lambda}_0))^2}{f(x; \boldsymbol{\lambda}_0)} dx > 1$.

Theorem 1. *If conditions (A1)-(A7) of Andrews et al. (2006) and assumptions (I)-(III) hold, there exists a sequence of local maximizers $\hat{\boldsymbol{\theta}}_{ML} = [\hat{\boldsymbol{\phi}}', \hat{\boldsymbol{\varphi}}', \hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\lambda}}', \hat{\sigma}]'$ of $L_T(\boldsymbol{\theta})$ in (11) such that*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \text{diag}(\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Omega}^{-1})).$$

3.2 Computing the Covariance Matrix

The covariance matrix $\boldsymbol{\Sigma}$ is of interest when one wants to compute (approximate) standard errors of the parameters for inference (e.g., confidence levels, hypothesis testing). A conventional estimator is based on the Hessian of the log-likelihood but nonlinear optimization of this function

often involves complicated gradient based numerical methods. As these procedures are relatively unstable in certain settings, we provide an alternative way to compute the standard errors of the parameters of both the autoregressive and exogenous variables for Student's t -MLE.

Similar to Hecq et al. (2016), we characterize the asymptotic distribution of the Student's t -MLE estimated parameters in the finite variance framework. If $\nu > 2$, the MLE is $\sqrt{n}$ -consistent and asymptotically normal. Define the $(n \times 1)$ series $\mathbf{u} \equiv \mathbf{U}_t^* = [u_{r+1}, \dots, u_{T-s}]'$ up to $\mathbf{U}_{t+s}^* = [u_{r+s+1}, \dots, u_T]'$, $\mathbf{V}_{t-r}^* = [v_1, \dots, v_{T-p}]'$ up to $\mathbf{v} \equiv \mathbf{V}_t^* = [v_{r+1}, \dots, v_{T-s}]'$, $\mathbf{X}_{i,t} = [x_{i,r+1}, \dots, x_{i,T-s}]'$ and $\boldsymbol{\varepsilon} = [\varepsilon_{r+1}, \dots, \varepsilon_{T-s}]'$. We construct $\mathbf{Z} = [\mathbf{U}_{t+1}^*, \dots, \mathbf{U}_{t+s}^*, \mathbf{X}_{1,t}, \dots, \mathbf{X}_{q,t}]$ and similarly $\mathbf{Q} = [\mathbf{V}_{t-1}^*, \dots, \mathbf{V}_{t-r}^*, \mathbf{X}_{1,t}, \dots, \mathbf{X}_{q,t}]$, which are of dimensions $(n \times (s+q))$ and $(n \times (r+q))$ respectively. Using this notation, we can write the autoregressions defined in (4) and (5) in matrix notation by $\mathbf{u} = \mathbf{Z}\boldsymbol{\zeta} + \boldsymbol{\varepsilon}$ and $\mathbf{v} = \mathbf{Q}\boldsymbol{\xi} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\zeta} = [\boldsymbol{\varphi}', \boldsymbol{\beta}']' \in \mathbb{R}^{s+q}$ and $\boldsymbol{\xi} = [\boldsymbol{\phi}', \boldsymbol{\beta}']' \in \mathbb{R}^{r+q}$. Then, conditional on the unobserved causal and noncausal components discussed above, it can be shown that in the case of an MARX(r, s, q) model

$$\sqrt{n}(\hat{\boldsymbol{\zeta}}_{ML} - \boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{\nu+3}{\nu+1} \sigma^2 \boldsymbol{\Upsilon}_{\phi}^{-1}\right), \quad (12)$$

$$\sqrt{n}(\hat{\boldsymbol{\xi}}_{ML} - \boldsymbol{\xi}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{\nu+3}{\nu+1} \sigma^2 \boldsymbol{\Upsilon}_{\varphi}^{-1}\right), \quad (13)$$

holds. We use the notation $\boldsymbol{\Upsilon}_{\varphi} = \mathbb{E}[\mathbf{Q}'\mathbf{Q}]$ and $\boldsymbol{\Upsilon}_{\phi} = \mathbb{E}[\mathbf{Z}'\mathbf{Z}]$, where φ and ϕ signify the relation between the unobserved values $\mathbf{u}, \mathbf{v}$ and $\mathbf{y}$ as defined in (4)-(5). These quantities can be estimated consistently by $(1/n) \sum_{i=1}^n \mathbf{Q}_i' \mathbf{Q}_i$ and $(1/n) \sum_{i=1}^n \mathbf{Z}_i' \mathbf{Z}_i$, where $\mathbf{Q}_i$ [resp. $\mathbf{Z}_i$] denotes the i th row of the matrix $\mathbf{Q}$ [resp. $\mathbf{Z}$]. For large ν , i.e., $\nu \rightarrow \infty$, $L_T(\boldsymbol{\theta})$ approaches the Gaussian (log)-likelihood, and the model parameters cannot be consistently estimated anymore.

Using the definitions above, results similar to Hecq et al. (2016) can be derived for the Least Absolute Deviations estimator. For MAR models, it can serve as initial estimator for $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$ (Lanne and Saikkonen, 2011) and outperforms Student's t -MLE in certain cases (Hecq et al., 2016). However, this density does not satisfy the regularity conditions of Andrews et al. (2006).

3.3 Forecasting

The MARX model might be preferred over the MAR model if the goal is to understand economic relationships, as the exogenous regressors are directly present in the regression equation. This offers the possibility to perform inference on coefficients linking these variables. By slightly adapting the methodology of Lanne et al. (2012a), we can also compute forecasts for the MARX model. The difference lies in the fact that forecasting of the noncausal component does not only involve ε_t but also $\mathbf{X}_t$. That is, the mean-square sense optimal one-step ahead forecast is given by

$$\mathbb{E}(y_{T+1}|\Omega_T) = \phi_1 y_T + \dots + \phi_r y_{T-r+1} + \mathbb{E} \left(\sum_{j=0}^{\infty} \delta_j z_{t+j+1} \middle| \Omega_T \right), \quad (14)$$

where z_{t+j} is defined as in (2). In practice, the last term (i.e. the infinite sum) is truncated by a sufficiently large number M to obtain a close approximation. The forecasting method can be described as follows. From the conditional distribution $(z_{T+1}, \dots, z_{T+M} | y_1, \dots, y_T)$ we simulate N mutually independent realizations of $(z_{T+1}, \dots, z_{T+M})$ which are plugged into (14) to obtain the one-step ahead forecast y_{T+1} .¹¹ We assume that M future values of $\mathbf{X}_t$ are known, such that they enter the density function as a constant. Abandoning this assumption requires a further extension of the forecasting procedure, as one needs the joint density of $(z_{T+1}, \dots, z_{T+M})$. The evaluation of the distribution of z_t is outside the scope of this paper. An alternative forecasting procedure for noncausal models is the sample-based method proposed by Gouriéroux and Jasiak (2016). An extensive comparison of the two methods can be found in Hecq and Voisin (2019).

In case of *ex post* forecasting, realized future values of $\mathbf{X}_t$ can be used to forecast y_t . Whenever we want to perform an *ex ante* forecasting exercise, one needs auxiliary equations to first forecast $\mathbf{X}_t$ and subsequently y_t . Note that assumption (I) ensures that the models for the exogenous variables do not depend on past, current and future values of y_t . We study both forecasting setups by simulation in the next section.

¹¹As proposed by Lanne et al. (2012a), we set $M = 50$ and $N = 10,000$ in this paper.

4 Simulation Study

By means of Monte Carlo simulations, we investigate four different cases of interest for MARX processes: (i) the performance of MLE, (ii) the model selection procedure, (iii) the identifiability under Gaussianity and (iv) forecasting. Each table in this simulation study reports results for 10,000 replications. For the cases (i) – (iii), the DGP is the MARX(1,1,1):

$$(1 - 0.3L)(1 - 0.5L^{-1})y_t - \beta_1 x_t = \varepsilon_t, \quad (15)$$

with the error term $\varepsilon_t \stackrel{iid}{\sim} t(3, 1)$, where $t(\nu, \sigma)$ denotes the Student's t distribution with ν degrees of freedom and scale parameter σ . For (i) and (ii), $\beta_1 = 0.3$, while the value of β_1 takes various values in (iii). Additionally, the process x_t will follow different specifications. The DGP for (iv) is detailed in that respective section.

4.1 Performance MLE for MARX

To assess the performance of the maximum likelihood estimator, four different specifications for x_t are considered: (i) $x_t \stackrel{iid}{\sim} t(5, 1)$, (ii) $x_t \stackrel{iid}{\sim} \mathcal{N}(0, 1)$, (iii) $x_t \stackrel{iid}{\sim} t(1, 1)$ (i.e. Standard Cauchy) and (iv) x_t follows an AR(1) process: $x_t = 0.6x_{1,t-1} + \epsilon_t$ where $\epsilon_t \stackrel{iid}{\sim} \mathcal{N}(0, 5)$. In every replication, processes are simulated and estimated by the ML estimator proposed in Section 3. Table 2 reports the mean and standard deviations of the estimated parameters by MLE over all 10,000 simulations. The scale parameter σ is fixed at 1 for both simulation and estimation.

Results are very similar for different specifications of x_t . The most noticeable difference is that the standard deviations of the parameters are larger for the first two cases especially when T is small. This can be due to the fact that both the $t(5, 1)$ and $\mathcal{N}(0, 1)$ distribution do not generate large outliers in x_t , making it more difficult to disentangle their contribution to the series from that of lags and leads of y_t . The means of the estimated parameters also lie further away from the true value when compared to the other specifications, but are still close. The

T	Parameter	Specification for x_t							
		$x_t \stackrel{iid}{\sim} t(5, 1)$		$x_t \stackrel{iid}{\sim} \mathcal{N}(0, 1)$		$x_t \stackrel{iid}{\sim} t(1, 1)$		$x_t \sim \text{AR}(1)$	
		Mean	Std. dev	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
50	ϕ_1	0.307	0.167	0.316	0.174	0.298	0.079	0.293	0.079
	φ_1	0.468	0.162	0.462	0.170	0.493	0.071	0.491	0.066
	β_1	0.303	0.152	0.305	0.194	0.299	0.039	0.304	0.036
	ν	5.140	8.562	5.089	8.203	5.050	8.295	5.416	9.906
100	ϕ_1	0.301	0.104	0.302	0.109	0.299	0.039	0.295	0.052
	φ_1	0.488	0.099	0.486	0.104	0.498	0.035	0.497	0.043
	β_1	0.302	0.102	0.302	0.131	0.300	0.018	0.302	0.024
	ν	3.487	1.899	3.487	2.216	3.474	1.754	3.544	2.534
500	ϕ_1	0.299	0.037	0.299	0.038	0.300	0.010	0.299	0.022
	φ_1	0.498	0.037	0.499	0.035	0.500	0.009	0.499	0.018
	β_1	0.301	0.043	0.300	0.055	0.300	0.004	0.300	0.011
	ν	3.057	0.366	3.061	0.367	3.053	0.423	3.060	0.364
1000	ϕ_1	0.300	0.025	0.300	0.026	0.300	0.006	0.300	0.015
	φ_1	0.499	0.023	0.499	0.023	0.500	0.004	0.500	0.013
	β_1	0.300	0.030	0.300	0.039	0.300	0.002	0.300	0.007
	ν	3.031	0.255	3.029	0.254	3.023	0.322	3.026	0.254

Table 2: Finite sample properties of the ML estimator for an MARX(1,1,1) with $\varepsilon_t \stackrel{iid}{\sim} t(3, 1)$

most difficult parameter to estimate is ν , which has a large standard deviation for $T = 50$. For larger T , the standard deviations decrease rapidly. In all cases, the estimated mean of all parameters becomes more accurate and standard deviations decline as T grows large.¹²

4.2 Model Selection

We now evaluate the proposed model selection procedure in Section 3.3. To that end, we simulate (15) with $x_t \stackrel{iid}{\sim} t(2, 1)$. We limit our attention to the estimation of purely causal ARX($p^*, 1$) models by Gaussian MLE, where $p^* = 0, \dots, 4$. Table 3 shows the percentages with which AIC, BIC and HQ select a certain order p^* (true order equals 2). As comparison, the same exercise has been done on the MAR model. That is, we consider the same specification only without

¹²The same simulation study was performed for infinite variance cases, e.g., $\varepsilon_t \stackrel{iid}{\sim} t(2, 1)$. Similar to results in Hecq et al. (2016) for the MAR model, the simulation study suggests the fatter the tails of the error distribution, the more accurate the estimation for both the coefficients and the distributional parameters of the MARX model.

exogenous variable x_t , i.e. $\beta_1 = 0$. The corresponding frequencies for the MAR model can be found in Table 4.

	$T = 100$			$T = 200$			$T = 500$			$T = 1000$		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
0	0.3	0.3	0.3	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0
1	61.6	82.4	72.6	55.3	74.8	65.3	30.0	47.0	34.9	0.2	4.1	0.7
2	33.5	16.4	25.2	41.8	24.1	32.9	67.1	52.9	63.8	96.4	95.5	97.6
3	4.6	0.9	1.9	2.6	0.8	1.6	2.9	0.1	1.3	3.4	0.4	1.7
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 3: Frequency (%) with which the autoregressive order p is selected for the MARX model

	$T = 100$			$T = 200$			$T = 500$			$T = 1000$		
p	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	41.0	67.7	53.3	26.3	56.8	39.0	3.5	19.6	8.1	0.0	1.8	0.2
2	52.7	30.1	43.2	63.8	40.8	56.0	86.9	78.9	87.3	84.7	96.5	93.4
3	6.3	2.2	3.5	10.0	2.4	5.0	9.6	1.5	4.6	15.3	1.7	6.4
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 4: Frequency (%) with which the autoregressive order p is selected for the MAR model

We can see that all information criteria tend to underestimate the true lag order (especially BIC) in small samples. The performance improves when T grows larger but at $T = 500$ we still observe correct autoregressive order selection in only around 65% of the cases. Whereas it is well-known that information criteria might not perform optimally in finite samples (see e.g., Hurvich and Tsai, 1989), the performance in the MARX setup for $T \in \{100, 200, 500\}$ is considerably worse when compared to the same exercise for MAR models. This stresses the usage of diagnostic tests to discover model fit improvements. For instance, in the empirical section we also test for autocorrelation (Box-Pierce and LM tests) after identifying the pseudo-causal model by information criteria. We increase the lag length p if necessary (see Fries and Zakoïan, 2019 for new Box-Pierce tests for MAR models).

Further, we suppose the correct total autoregressive order $p = 2$ is known and investigate

the selection of $\text{MARX}(r, s, 1)$ and $\text{MAR}(r, s)$ models with $r + s = 2$ based on the highest log-likelihood. We consider two cases for the MARX model: (i) shifting x_t along the lines of Section 2.2 and (ii) considering x_t at the same point in time for all models.¹³ In the first case, we compare models that are equivalent from a theoretical point of view, while the second case provides alternatives that are often empirically a better fit. In Table 5, we observe that the model selection procedure improves with sample size for all models considered. The MARX model with shifted x_t , denoted $\text{MARX}(s)$, performs best and shows extremely satisfactory results already at $T = 100$. The identification is likely to be established more swiftly as there are two types of information in the DGP: (i) the type of dynamics (i.e. causal, noncausal or mixed) and (ii) how x_t is linked with respect to y_t in time. The MAR and the MARX model without shift do not contain the second source of information, leading to more doubt between the true mixed or the misspecified noncausal specification. However, results are still extremely satisfactory.

	$T = 100$			$T = 200$			$T = 500$			$T = 1000$		
	(2,0)	(1,1)	(0,2)	(2,0)	(1,1)	(0,2)	(2,0)	(1,1)	(0,2)	(2,0)	(1,1)	(0,2)
$\text{MARX}(s)$	1.22	98.55	0.23	0.02	99.98	0.00	0.00	100.0	0.00	0.00	100.00	0.00
MARX	0.24	87.82	11.94	0.01	96.11	3.88	0.00	99.70	0.30	0.00	100.00	0.00
MAR	2.42	89.54	8.04	0.31	97.71	1.98	0.00	99.91	0.09	0.00	100.00	0.00

Table 5: Lag-lead order (r, s) selected by highest log-likelihood with fixed $p = 2$ for both MARX and MAR model (in %)

The model selection procedure is sensitive to the values of the parameters chosen in the DGP as well as the chosen error distribution (and how it deviates from Gaussianity). For example, consider the case in which $\phi_1 = 0.1$ and $\varphi_1 = 0.7$. Since the value of ϕ_1 is so low, first-order models are selected too often in the first step of the model selection procedure. As φ_1 is much larger than ϕ_1 , a noncausal model will be selected in the second step. If we consider $\phi_1 = 0.5$ and $\varphi_1 = 0.7$ instead, it is less likely that first-order models are selected too often in the first step. Hence, it is also more probable that the “correct” mixed causal-noncausal model is eventually

¹³For the sake of simplicity, we do not report the results of the model selection with all possible combinations as introduced in Section 2.3. These simulation results are available upon request.

chosen in the second step. This suggests the use of complementary analysis (e.g., bootstrap or cross-validation). Gouriéroux and Jasiak (2018) show that underestimating the lag order p is less probable using OLS compared to non-Gaussian AML.

4.3 Identifiability

We now study whether MARX processes can be identified using solely second-order properties. We simulate an MARX(1,1,1) process from (15), where we consider two different specifications for the exogenous regressor: (i) $x_t \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ and (ii) x_t follows a causal AR(1) process: $x_t = 0.6x_{1,t-1} + \epsilon_t$ where $\epsilon_t \stackrel{iid}{\sim} \mathcal{N}(0, 5)$. In the first case x_t exhibits no dynamics, while it does in the second specification. The coefficient β_1 takes several values. After the process is generated, three models are estimated by Gaussian MLE: MARX(2,0,1), MARX(1,1,1) and MARX(0,2,1). The selected model is the one that attains the highest log-likelihood. The results for the first case can be found in Table 6. We use Gaussian MLE to estimate the models due to the multiplicative structure of polynomials in the MARX(1,1,1) specification.

β_1	$T = 100$	$T = 200$	$T = 500$	$T = 1000$
0	41.27	41.72	27.28	33.35
0.3	69.37	84.51	94.63	99.56
0.5	89.20	97.32	99.70	99.97
0.8	98.17	99.59	99.98	99.99

Table 6: Frequency (%) with which the correct model is chosen

On the one hand, it can be observed that the correct model cannot be identified when $\beta_1 = 0$. Since the process reduces in this case to an MAR, it is expected that second-order properties cannot distinguish backward- and forward-looking behavior. Hence, we would expect to pick all three models with equal probability. On the other hand, we observe that identification of the correct model is more accurate both when T increases and when the exogenous regressor becomes more important, i.e. β_1 increases in absolute value. When x_t follows an AR(1) process, results are very similar. Frequencies are very close to 100% already when $\beta_1 = 0.3$ for all T .

The identification of MARX models using Gaussian MLE is highly dependent on two factors: (i) the value of the coefficients of the exogenous regressors and (ii) the type of dynamics the exogenous regressors exhibit.

As a final check, we briefly comment on identifiability based on different estimation procedures. To that end, we consider the following DGP:

$$y_t = 0.6y_{t+1} + 0.5x_t + \varepsilon_t, \quad (16)$$

with $\varepsilon_t \stackrel{iid}{\sim} t(3, 1)$ and x_t standard normal. We estimate the “true” model in (16) and the pseudo-causal model (with shift in x_t) by OLS, Gaussian MLE and t -MLE. The model with the highest log-likelihood is chosen to be the final model. Table 7 shows the frequency with which the true noncausal model is chosen. Two main conclusions can be drawn: (i) t -MLE outperforms OLS as well as Gaussian MLE and (ii) frequencies are above 85% already for $T = 100$ for all estimation procedures. Results are even better when x_t exhibits dynamics.¹⁴ It has to be mentioned that simulation results might differ depending on parameter values chosen. To that end, we chose rather moderate values for both the autoregressive and the exogenous regressor’s coefficient.

	$T = 100$	$T = 200$	$T = 500$	$T = 1000$
OLS	85.41	94.74	99.26	99.87
Gaussian MLE	85.57	94.81	99.26	99.88
t -MLE	98.96	99.98	100.00	100.00

Table 7: Frequency (%) with which the correct model is chosen

A closer inspection of the misspecified models reveals that they behave differently based on among others (i) the value chosen for φ_1 , (ii) the value chosen for β_1 , (iii) the type of process x_t (e.g. static, dynamic) and (iv) the distribution of x_t and ε_t . The exact link between the true DGP and the corresponding misspecified models can be studied in the spirit of Gouriéroux and Jasiak (2018), but this is outside the scope of this paper.

¹⁴Simulation results where x_t follows an AR(1) process are available upon request.

4.4 Forecasting

In this section, the DGP is the MARX(0,1,1) where the exogenous regressor x_t is assumed to follow a causal AR(1) process:

$$y_t = 0.7y_{t+1} + 0.8x_t + \varepsilon_t, \quad (17)$$

$$\text{with } x_t = \beta x_{t-1} + w_t, \quad (18)$$

with $\varepsilon_t \stackrel{iid}{\sim} t(3, 1)$. We examine a low persistence ($\beta = 0.3$) and a high persistence ($\beta = 0.8$) case. For both scenarios, we also consider two different distributions for the error term w_t , i.e. $\mathcal{N}(0, 1)$ and $t(3, 1)$, in the auxiliary regression equation. Two models are considered for forecasting y_t : the MARX(0,1,1) and MAR(0,1). From (17), we use $T = 200$ observations to estimate the model and forecast $h = 8$ observations using one-step-ahead forecasts. To make a broad comparison of the forecasting performance of MARX and MAR models, we consider several setups.

For the MARX model, we examine an *ex post* and *ex ante* way to forecast. For the first case, we use realized future values of x_t , while we forecast the exogenous regressor directly as a causal AR(1) under known parameter value for the second case. We denote these specifications MARX1¹⁵ and MARX2 respectively. For the MAR model, we also study two setups. In the first case, we fit an MAR model to y_t and produce forecasts based on these parameter estimates. In the second case, forecasts are computed according to (17) with the only difference that we do not take the exogenous regressor into account. In this way, we want to assess the effect of an omitted variable on the forecasting performance. These two models are denoted MAR1 and MAR2 respectively.

Table 8 displays the average root mean squared forecast error (RMSFE) over all replications for the four setups considered. In all cases, MARX1 has the lowest RMSFE and performs best.

¹⁵For MARX1, we keep an additional $M = 50$ future values of x_t to forecast using the MARX(0,1,1).

	low persistence ($\beta = 0.3$)		high persistence ($\beta = 0.8$)	
	$w_t \stackrel{iid}{\sim} \mathcal{N}(0, 1)$	$w_t \stackrel{iid}{\sim} t(3, 1)$	$w_t \stackrel{iid}{\sim} \mathcal{N}(0, 1)$	$w_t \stackrel{iid}{\sim} t(3, 1)$
MARX1	2.040	1.923	1.983	1.965
MARX2	2.386	2.766	2.540	4.626
MAR1	2.441	2.842	2.543	4.703
MAR2	2.400	2.798	2.499	4.993

Table 8: RMSFE for different model specifications

This result is not surprising, as true future information is used to compute forecasts of y_t .¹⁶ The RMSFE for the remaining three models follow on a respectable distance. We observe that the MARX2 provides forecasting advantages whenever x_t contributes considerably to the formation of y_t , i.e. when w_t follows a fat-tailed distribution. Note that the benefits are larger in the high persistence case, since it takes longer before forecasts for x_t return to their unconditional mean. MAR1 outperforms MAR2 in these instances, as the predictable effect of the exogenous regressor is absorbed in the noncausal parameter. For the cases where x_t contributes only minimally to the construction of y_t , both MAR forecasts are often even outperforming the MARX2. As the RMSFE of the MARX2, MAR1 and MAR2 lie relatively close to each other, there is no clear evidence favoring the MAR or MARX model.

5 Empirical Application

5.1 The Data

We consider non seasonally adjusted monthly commodity prices $CP_{i,t}$ from 1980M1 to 2019M4, i.e., 472 observations for $i = 1, \dots, 5$ indexes released by the IMF.¹⁷ These are benchmark prices determined by the largest exporter of a given commodity and are thus representative of the global market. IMF releases many different individual commodity prices but we only focus on the following indexes for this study: (i) BEVE: Beverage Price Index, (ii) COP: Copper

¹⁶We do however note that ex post forecasting is very common in the literature on ARX models.

¹⁷IMF Primary Commodity Prices, see <http://www.imf.org/external/np/res/commod/index.aspx>.

Price Index (*iii*) NICK: Nickel Price Index, (*iv*) OIL: Crude Oil Price index and (*v*) RAWM: Agricultural Raw Materials Index. These five commodities serve as our variables of interest. As exogenous regressors, we consider the trade weighted U.S. dollar index (S_t) and the industrial production index (IP_t). These series are taken from the Federal Reserve Bank of St. Louis database.¹⁸ We justify strict exogeneity of the U.S. industrial production index and the U.S. dollar index by the small size of the U.S. economy on trade for almost all of these commodities. One potential problem in making this statement is that the classifications of the commodity indices do not exactly match those of trade statistics. However, we get a very good fit in almost all cases. For example, regarding RAWM, for 2016, the U.S. share of raw material exports and imports were 9.42% and 8.43% respectively. For OIL, in 2017, U.S. exports represented a mere 2.6% of the total.

[FIGURE 3 AROUND HERE]

The method to detrend series before identifying MAR models is an ongoing debate. Hencic and Gouriéroux (2014) fit a cubic deterministic trend to bitcoin data. We rely on conventional unit root analysis. Using ADF tests, we do not reject a unit root at the 5% significance level in each series. We consequently work with monthly growth rates (in %): $\Delta cp_{i,t} \equiv 100(1 - L) \ln CP_{i,t}$, $\Delta ip_t \equiv 100(1 - L) \ln IP_t$ and $\Delta s_t \equiv 100(1 - L) \ln S_t$, which are plotted in Figure 2

5.2 Identification of Pseudo-Causal Models

Let $y_{i,t} = \Delta cp_{i,t}$ and $\mathbf{X}_t = [\Delta s_t, \Delta ip_t]'$. As outlined in Section 2.3, the first step of our modelling strategy consists of identifying the pseudo-causal model for the dependent variable $y_{i,t}$. Since we do not demean the variables, we include an intercept in all models. The first part of Table 9 reports the lag order chosen by HQ after estimating ARDL models up to order $p_{max} = 8$ (both for $y_{i,t}$ and $\mathbf{X}_t$). The reason we rely on HQ is because the simulation results of Section 4.2 show

¹⁸Series name TWEXBMTH and INDPRO at <https://fred.stlouisfed.org>.

that BIC tends to underestimate the true order more often than AIC and HQ. Additionally, we check correlograms and perform LM tests for the null of no autocorrelation on the residuals. These results are in line with the lag order chosen by HQ. For each of the commodities, there are no lags selected for the exogenous regressors.¹⁹ Moreover, a simple regression between the two exogenous variables in $\mathbf{X}_t$ shows that we do not reject the null of linear independence.

The second part of Table 9 reports the estimation results with heteroskedasticity-consistent standard errors (HCSE) in brackets. We observe that commodity prices depend on their own lags as well as on the exchange rate and industrial production (except for OIL and BEVE) at the 5% significance level. The highest negative effect of the growth rate of the exchange rate is on the COP, OIL and NICK index. This result seems plausible, as these products heavily depend on exports and are thus more negatively influenced by an increase of the US dollar. The last rows of Table 9 report adjusted R^2 as well as the p -values of the Jarque-Bera test for normality and the LM ARCH[1,2] test for no autocorrelation up to order two in the squared residuals. Both tests follow a χ^2 distribution with two degrees of freedom under the null. It is observed that we reject the null of normality in every equation (except for RAWM).

5.3 Identification and Estimation of MARX

The first step of the model selection procedure shows that four out of five commodities have an autoregressive order equal to one. We characterize the remaining steps of the procedure presented in Section 2.3 for these series. Table 10 displays the values of the log-likelihood for four different models. The first two rows represent the models that are theoretically equivalent using the ARDL as starting point; the last two rows show equivalent models with the noncausal model as initial choice. Note that there are no remaining models to be tested as specified in Step 4 of the algorithm.

¹⁹This step can be done using (for instance) EViews ARDL estimation features. The same step can be performed in the R package **MARX** by allowing for lags in $\mathbf{X}_t$ and comparing information criteria.

	Commodities				
	BEVE	COP	NICK	OIL	RAWM
PHQ	1	1	1	1	2
c	0.001 (0.002)	0.004 (0.003)	0.003 (0.004)	0.004 (0.004)	0.002 (0.001)
a_1	0.270 (0.069)	0.263 (0.058)	0.299 (0.044)	0.274 (0.063)	0.194 (0.046)
a_2					0.169 (0.045)
b_1	-0.554 (0.180)	-1.652 (0.280)	-1.564 (0.333)	-1.613 (0.321)	-0.524 (0.109)
b_2	0.571 (0.289)	0.849 (0.402)	1.267 (0.488)	0.965 (0.534)	0.240 (0.192)
$\bar{R}^2$	0.104	0.236	0.176	0.165	0.162
JB	< 0.001	< 0.001	< 0.001	< 0.001	0.069
LM ARCH[1,2]	< 0.001	0.009	0.871	< 0.001	0.008

Table 9: Estimation results for pseudo-causal models by OLS, HCSE in ($\cdot$)

Candidate models	BEVE	OIL	NICK	COP
MARX(1,0,2) with $\mathbf{X}_t$	776.623	563.314	585.320*	732.792*
MARX(0,1,2) with $\mathbf{X}_{t+1}$	781.880	549.099	566.490	701.597
MARX(0,1,2) with $\mathbf{X}_t$	784.588*	565.910*	577.364	723.192
MARX(1,0,2) with $\mathbf{X}_{t-1}$	772.447	546.849	570.342	706.584

Table 10: Values for the log-likelihood of different models, * indicates the highest log-likelihood for each commodity

For BEVE and OIL a purely noncausal model with contemporaneous exogenous regressors is selected, while NICK and COP admit a purely causal representation. Note that, since RAWM has a different autoregressive order than the other commodities, it is excluded from Table 10, as the set of candidate models is larger. More specifically, since $p = 2$, we follow the procedure as presented in Example 2 to find the model with the highest log-likelihood. The selected model is the MARX(0,2,2) with $\mathbf{X}_t$ contemporaneous with a value of the log-likelihood equal to 1039.573. However, it should be noted that the purely causal specification is very close with a value of 1039.312. As the null of normality cannot be rejected on the residuals of the RAWM series, an identification problem might occur. It is well-known that, under a Gaussian error term, one

cannot distinguish between MAR processes. In the RAWM case, the presence of exogenous regressors does not create a clear asymmetry in different time directions, which also complicates identification for the MARX.

Table 11 reports the final results for each commodity; standard errors (SE) in brackets. Distributions are leptokurtic: the smallest value for the degrees of freedom parameter is $\hat{\nu} = 3.833$ for COP, the largest value is $\hat{\nu} = 13.186$ for RAWM. This result is in line with the result of the JB test. The negative impact of the exchange rate is more pronounced for OIL and the smallest value is for the BEVE commodity index. The coefficient of the industrial production index is not significantly different from zero in the OIL and BEVE equations. For all five commodity series we do not reject the null of no autocorrelation, but we still observe non-normality and heteroskedasticity using standard diagnostic tests on the residuals of the identified MARX models. Note that tests for homoskedasticity proposed by Gouriéroux and Zakoïan (2016) and Cavaliere et al. (2018) assume that the true model is noncausal with *iid* Cauchy error term, which is not the case here, as the most series are heteroskedastic both in direct and reverse time.

Lanne and Saikkonen (2011) claim that the errors in mixed causal-noncausal models contain effects of omitted variables that are predictable by the considered series. Suppose the variables $\mathbf{X}_t$ are indeed part of the DGP of the commodity prices $y_{i,t}$ but we fail to include them in our model. The identified models can be found in Table 12. We find that, in all cases except for RAWM, the causal and noncausal orders of the MAR and MARX are chosen to be the same. However, the estimated values of both the causal and noncausal coefficients are larger in absolute value for the MAR model. Additionally, we observe that the degrees of freedom parameters are lower for the MAR models. Small differences would indicate that the MARX models involve only minor benefits compared to the corresponding MAR model. A special case is the RAWM series, for which the degrees of freedom parameter is estimated to be much lower in the MAR specification ($\hat{\nu} = 6.830$ compared to $\hat{\nu} = 13.114$ for the MARX). Values of the log-likelihood

	Commodities				
	BEVE	COP	NICK	OIL	RAWM
	MARX(0,1,2)	MARX(1,0,2)	MARX(1,0,2)	MARX(0,1,2)	MARX(0,2,2)
α	-0.118 (0.190)	0.398 (0.211)	0.045 (0.300)	0.771 (0.394)	0.162 (0.119)
ϕ_1		0.278 (0.034)	0.296 (0.037)		
φ_1	0.267 (0.037)			0.231 (0.037)	0.202 (0.043)
φ_2					0.155 (0.043)
β_1	-0.335 (0.144)	-1.295 (0.160)	-1.290 (0.227)	-1.526 (0.230)	-0.496 (0.090)
β_2	0.222 (0.279)	0.783 (0.310)	1.451 (0.441)	0.364 (0.447)	0.287 (0.174)
ν	3.965 (0.761)	3.831 (0.767)	5.016 (1.048)	4.058 (0.812)	13.114 (7.240)
σ	3.471 (0.202)	3.840 (0.238)	5.627 (0.295)	5.573 (0.330)	2.393 (0.128)

Table 11: Estimation results for MARX models by Student's t -MLE, SE in $(\cdot)$

for the different MAR models lie relatively far from another, which allows clear identification. The MARX however shows that the fat-tailedness of the series is mostly due to the omission of explanatory variables.

A few additional observations are in place. First, the MARX model has the advantage that it recovers the possibility to identify using second-order properties. As explained in Section 2.4 and 4.3, estimation methods such as OLS and Gaussian MLE can be used to confirm results. However, there, we also highlight that successful identification depends on factors such as the type of process generating the exogenous regressors as well as the size of the contribution of these regressors for determining y_t (i.e. the value of their coefficients). Second, we discussed in Section 2 that the degrees of the causal and noncausal polynomials might increase depending on the dynamics of the exogenous regressors. As the exchange rate closely resembles a white noise series, it is not surprising that the orders did not increase in this empirical application.

	Commodities				
	BEVE	COP	NICK	OIL	RAWM
	MAR(0,1)	MAR(1,0)	MAR(1,0)	MAR(0,1)	MAR(1,1)
α	-0.178 (0.190)	0.176 (0.225)	-0.144 (0.310)	0.389 (0.318)	0.058 (0.117)
ϕ_1		0.329 (0.036)	0.338 (0.038)		-0.334 (0.041)
φ_1	0.284 (0.037)			0.287 (0.038)	0.561 (0.036)
ν	3.887 (0.695)	3.824 (0.708)	4.889 (1.024)	4.380 (0.913)	6.830 (2.083)
σ	3.465 (0.201)	4.110 (0.239)	5.805 (0.310)	5.893 (0.340)	2.261 (0.126)

Table 12: Estimation results for MAR models by Student’s t -MLE, SE in $(\cdot)$

We do not claim that these models capture all the dynamics in the series. For example, it could be the case that one needs to include more exogenous regressors or that a system of equations (e.g. the noncausal VAR) is a more adequate modelling approach. However, we want to point out that, since the inclusion of exogenous regressors clearly affects the autoregressive parameters, one should be cautious when interpreting the values of the structural parameters in (10). The main debate in the literature evolves around the importance of forward-looking compared to backward-looking behavior. Results in this empirical application suggest that this comparison might be distorted whenever relevant exogenous regressors are omitted.

6 Conclusion

This paper proposes to estimate mixed causal-noncausal models including exogenous regressors by non-Gaussian MLE. A detailed description of the method to estimate and select MARX models is given, in which the implications of explanatory variables is highlighted. The proposed estimation of standard errors is robust to computational overflows. We find that the presence of exogenous regressors has the feature that one can potentially discriminate between causal,

noncausal and mixed causal-noncausal specifications based on solely second-order properties. Monte Carlo simulation studies are conducted to assess the performance of the proposed estimation method and model selection procedure in finite samples. Additionally, the forecasting performance of MARX models is compared with MAR models in several setups, which suggests that the former is slightly favored in case of ex post forecasting.

We provide an empirical study on the relation between commodity prices, the exchange rate and the industrial production index. We identify noncausal models for BEVE, OIL and RAWM, and causal models for NICK and COP based on maximizing the log-likelihood. A comparison is made with the selection of MAR models. Whereas the same causal and noncausal orders are selected in most cases, we observe that the estimated autoregressive parameters are larger in absolute value, while the estimates for the degrees of freedom parameter are lower. This suggests that the MAR model compensates for the omission of relevant regressors through these channels.

References

- Alessi, L., Barigozzi, M. & Capasso, M. (2011). Non-Fundamentality in Structural Econometric Models: A Review. *International Statistical Review*, 79, 1.
- Andrews, B., Breidt, F. & Davis, R. (2006). Maximum Likelihood Estimation For All-Pass Time Series Models. *Journal of Multivariate Analysis*, 97, 1638-1659.
- Breidt, F., Davis, R., Lii, K. & Rosenblatt, M. (1991). Maximum Likelihood Estimation for Noncausal Autoregressive Processes. *Journal of Multivariate Analysis*, 36, 175-198.
- Brockwell, P. & Davis, R. (1991). *Time Series: Theory and Methods*. Springer-Verlag New York, Second Edition.
- Cavaliere, G., Nielsen, H. B. & Rahbek, A. (2018). Bootstrapping Noncausal Autoregressions:

- With Applications to Explosive Bubble Modeling. *Journal of Business and Economic Statistics*, doi:10.1080/07350015.2018.1448830.
- Cubadda, G., Hecq, A. & Telg, S. (2019). Serial Correlation Common Noncausal Features. *Oxford Bulletin of Economics and Statistics*, 81(3), 697-715.
- Fries, S. & J.-M. Zakoian (2019). Mixed Causal-Noncausal AR Processes and the Modelling of Explosive Bubbles. *Econometric Theory*, 1-37, doi:10.1017/S0266466618000452
- Gouriéroux, C. & Jasiak, J. (2016). Filtering, Prediction and Simulation Methods in Noncausal Processes. *Journal of Time Series Analysis*, doi: 10111/jtsa.12165.
- Gouriéroux, C. & Jasiak, J. (2018). Misspecification of noncausal order in autoregressive processes. *Journal of Econometrics*, 205(1), 226–248.
- Gouriéroux, C. & Zakoïan, J.M. (2016). Local Explosion Modelling by Noncausal Process. *Journal of the Royal Statistical Society, Series B*, doi:10.1111/rssb.12193.
- Hecq, A., Lieb, L. & Telg, S. (2016). Identification of Mixed Causal-Noncausal Models in Finite Samples. *Annals of Economics and Statistics*, 123/124, 307-331.
- Hecq, A., Lieb, L. & Telg, S. (2017a). Simulation, Estimation and Selection of Mixed Causal-Noncausal Autoregressive Models: The MARX Package. Working Paper, available at SSRN: <https://ssrn.com/abstract=3015797>.
- Hecq, A., Telg, S. & Lieb, L. (2017b). Do Seasonal Adjustments Induce Noncausal Dynamics in Inflation Rates?. *Econometrics*, 5(4), 48.
- Hecq, A. & Voisin E. (2019). Forecasting Bubbles with Mixed Causal-Noncausal Autoregressive Models. MPRA Working Paper, University Library of Munich.

- Hencic, A. & Gouriéroux C. (2014). Noncausal Autoregressive Model in Application to Bitcoin/USD Exchange Rate. *Econometrics of Risk, Series: Studies in Computational Intelligence*, Springer International Publishing, 17-40.
- Hurvich, M. & Tsai, C.L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76, 297-307.
- Issler, J.V, Rodrigues, C. and Burjack, R. (2014). Using Common Features to Understand the Behavior of Metal-Commodity Prices and Forecast Them at Different Horizons. *Journal of International Money and Finance*, 42, 310-335.
- Karapanagiotidis, P. (2014). Dynamic Modeling of Commodity Futures Prices. MPRA Paper 56805, University Library of Munich, Germany.
- Lanne, M., Luoto J. & Saikkonen, P. (2012a). Optimal Forecasting of Noncausal Autoregressive Time Series. *International Journal of Forecasting*, 28, 623-631.
- Lanne, M., Nyberg, H. & Saarinen, E. (2012b). Does Noncausality Help in Forecasting Economic Time Series? *Economics Bulletin*, 32(4), 2849-2859.
- Lanne, M. & Luoto, J. (2013). Autoregression-Based Estimation of the New Keynesian Phillips Curve. *Journal of Economic Dynamics & Control*, 37, 561-570.
- Lanne, M. & Saikkonen, P. (2011). Noncausal Autoregressions for Economic Time Series. *Journal of Time Series Econometrics*, 3(3), 1-32.
- Lof, M. (2013). Noncausality and asset pricing. *Studies in Nonlinear Dynamics and Econometrics*, 17, 211-220.
- Lof, M. & Nyberg, H. (2017). Noncausality and the Commodity Currency Hypothesis. *Energy Economics*, 65, 424-433.

Appendix A - Definitions for Lemma 2

We first consider the score of $\boldsymbol{\theta}$ evaluated at true parameter values. Define $\mathbf{V}_{t-1} = [v_{t-1}, \dots, v_{t-r}]'$ and $\mathbf{U}_{t+1} = [u_{t+1}, \dots, u_{t+s}]'$, where u_t and v_t are defined in terms of true parameter values, i.e., $u_t = \sum_{j=0}^{\infty} \delta_{0j} (\sum_{i=1}^q \beta_{0i} x_{i,t+j} + \varepsilon_{t+j})$ and $v_t = \sum_{j=0}^{\infty} \alpha_{0j} (\sum_{i=1}^q \beta_{0i} x_{i,t-j} + \varepsilon_{t-j})$. By direct differentiation of (11), we obtain:

$$\frac{\partial}{\partial \boldsymbol{\phi}} g_t(\boldsymbol{\theta}_0) = -\frac{f'(\sigma_0^{-1} \varepsilon_t; \boldsymbol{\lambda}_0)}{\sigma_0 f(\sigma_0^{-1} \varepsilon_t; \boldsymbol{\lambda}_0)} \mathbf{V}_{t-1}, \quad \frac{\partial}{\partial \boldsymbol{\varphi}} g_t(\boldsymbol{\theta}_0) = -\frac{f'(\sigma_0^{-1} \varepsilon_t; \boldsymbol{\lambda}_0)}{\sigma_0 f(\sigma_0^{-1} \varepsilon_t; \boldsymbol{\lambda}_0)} \mathbf{U}_{t+1},$$

and

$$\frac{\partial}{\partial \boldsymbol{\beta}} g_t(\boldsymbol{\theta}_0) = -\frac{f'(\sigma_0^{-1} \varepsilon_t; \boldsymbol{\lambda}_0)}{\sigma_0 f(\sigma_0^{-1} \varepsilon_t; \boldsymbol{\lambda}_0)} \mathbf{X}_t,$$

where $f'(x; \boldsymbol{\lambda}) = \partial f(x; \boldsymbol{\lambda}) / \partial x$ and use has been made of the fact that $\varphi_0(L^{-1})u_t - \boldsymbol{\beta}'_0 \mathbf{X}_t = \varepsilon_t = \phi_0(L)v_t - \boldsymbol{\beta}'_0 \mathbf{X}_t$. Similarly, for the distributional parameters:

$$\frac{\partial}{\partial \sigma} g_t(\boldsymbol{\theta}_0) = -\sigma_0^2 \left(\frac{f'(\sigma_0^{-1} \varepsilon_t; \boldsymbol{\lambda}_0)}{f(\sigma_0^{-1} \varepsilon_t; \boldsymbol{\lambda}_0)} + \sigma_0 \right), \quad \frac{\partial}{\partial \boldsymbol{\lambda}} g_t(\boldsymbol{\theta}_0) = \frac{1}{f(\sigma_0^{-1} \varepsilon_t; \boldsymbol{\lambda}_0)} \frac{\partial}{\partial \boldsymbol{\lambda}} f(\sigma_0^{-1} \varepsilon_t; \boldsymbol{\lambda}_0).$$

Lemma 2 presents the asymptotic distribution of the score vector. Define $\mathcal{J} \equiv \int \frac{(f'(x; \boldsymbol{\lambda}_0))^2}{f(x; \boldsymbol{\lambda}_0)} dx > 1$ and $\mathcal{I} \equiv \int x^2 \frac{(f'(x; \boldsymbol{\lambda}_0))^2}{f(x; \boldsymbol{\lambda}_0)} dx - 1$, where the first inequality follows from Remark 2 in Andrews et al. (2006). Furthermore set

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} \\ \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33} \end{bmatrix}.$$

The matrix $\boldsymbol{\Sigma}$ is symmetric and has the matrices $\boldsymbol{\Sigma}_{11} = \sigma_0^{-2} \mathcal{J} \boldsymbol{\Gamma}_V$, $\boldsymbol{\Sigma}_{22} = \sigma_0^{-2} \mathcal{J} \boldsymbol{\Gamma}_U$ and $\boldsymbol{\Sigma}_{33} = \sigma_0^{-2} \mathcal{J} \boldsymbol{\Gamma}_X$ on the diagonal, where $\boldsymbol{\Gamma}_V$ and $\boldsymbol{\Gamma}_U$ are the autocovariance matrices of $\mathbf{V}_{t-1}$ and $\mathbf{U}_{t+1}$ respectively. $\boldsymbol{\Gamma}_X$ is the cross-covariance matrix of the q processes in $\mathbf{X}_t$ which is diagonal under the assumption of linear independence between processes in $\mathbf{X}_t$. $\boldsymbol{\Sigma}_{12}$ is a $(r \times s)$ matrix where the

(i, j) th element equals: $\sum_{t=0}^{\infty} \alpha_t \delta_{t+i-j} + \tilde{\mathcal{J}} \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \alpha_a \delta_b \sum_{m=1}^q \beta_m^2 \gamma_{x_m}(i+j+a+b)$. The Σ_{13} matrix has size $(r \times q)$ with the (i, j) th element $\beta_j \sigma^{-2} \mathcal{J} \sum_{a=0}^{\infty} \alpha_a \gamma_{x_j}(i+a)$, while this element equals $\beta_j \sigma^{-2} \mathcal{J} \sum_{b=0}^{\infty} \delta_b \gamma_{x_j}(i+b)$ for Σ_{23} . Note that the summands involve only i and not j , as only the former denotes the lag or lead considered for v_t and u_t respectively. In contrast, j ($= 1, \dots, q$) runs over all exogenous variables $x_t, \dots, x_{q,t}$. Finally the $(d+1) \times (d+1)$ matrix Ω is defined as

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

where $\Omega_{\lambda\lambda} \equiv \int \frac{1}{f(x; \lambda_0)} \left(\frac{\partial}{\partial \lambda} f(x; \lambda) \right) \left(\frac{\partial}{\partial \lambda} f(x; \lambda) \right)' dx$, $\omega_{\lambda\sigma} \equiv -\sigma_0 \int x \frac{f'(x; \lambda_0)}{f(x; \lambda_0)} \frac{\partial}{\partial \lambda} f(x; \lambda_0) dx = \omega'_{\sigma\lambda}$ and $\omega_{\sigma}^2 \equiv \omega_0^{-2} \mathcal{I}$.

Graphs

Figure 1: A simulated MARX process (left) and the same process without exogenous variable (right)

Figure 2: Growth rates of commodity prices, exchange rate and industrial production index

**ONLINE APPENDIX FOR:
MIXED CAUSAL-NONCAUSAL AUTOREGRESSIONS
WITH EXOGENOUS REGRESSORS**

Alain Hecq

*Maastricht University, School of Business and Economics, Department of Quantitative
Economics, P.O. Box 616, 6200 MD Maastricht, The Netherlands*
<a.hecq@maastrichtuniversity.nl>

João Victor Issler

*Graduate School of Economics - EPGE, Getulio Vargas Foundation, Praia de Botafogo
190s. 1100, Rio de Janeiro, RJ 22250-900, Brazil*
<Joao.Issler@fgv.br>

Sean Telg

*Vrije Universiteit Amsterdam, Department of Econometrics, De Boelelaan 1105,
1081 HV Amsterdam, The Netherlands*
<j.m.a.telg@vu.nl>

Abstract

This online appendix provides supplementary material for the paper “Mixed Causal-Noncausal Autoregressions with Exogenous Regressors”. The material consists of the proofs of the main results, an auxiliary proposition (with proof), the derivation of the approximate likelihood function and how the MARX model can arise from a transfer function model.

A.1 Proof of Lemma 2

Define $e_t \equiv \frac{f'_\sigma(\varepsilon_t; \boldsymbol{\lambda})}{f_\sigma(\varepsilon_t; \boldsymbol{\lambda})} \equiv \frac{f'(\varepsilon_t/\sigma; \boldsymbol{\lambda})}{\sigma f(\varepsilon_t/\sigma; \boldsymbol{\lambda})}$, $\tilde{\mathcal{J}} \equiv \sigma^{-2} \mathcal{J}$, $\tilde{\mathcal{I}} \equiv \sigma^{-2} \mathcal{I}$ and $n \equiv (T - p)$. Furthermore, let $x \equiv \varepsilon_t/\sigma$, then we have that

$$\mathbb{E}(e_t^2) = \mathbb{E} \left[\left(\frac{f'_\sigma(\varepsilon_t; \boldsymbol{\lambda})}{f_\sigma(\varepsilon_t; \boldsymbol{\lambda})} \right)^2 \right] = \int \left(\frac{f'_\sigma(\varepsilon_t; \boldsymbol{\lambda})}{f_\sigma(\varepsilon_t; \boldsymbol{\lambda})} \right)^2 f_\sigma(\varepsilon_t; \boldsymbol{\lambda}) d\varepsilon_t = \sigma^{-2} \int \frac{(f'(x; \boldsymbol{\lambda}))^2}{f(x; \boldsymbol{\lambda})} dx = \tilde{\mathcal{J}},$$

where we used the definitions of the density and $\mathcal{J}$. Also we have that

$$\mathbb{E}(e_t) = \mathbb{E} \left(\frac{f'_\sigma(\varepsilon_t; \boldsymbol{\lambda})}{f_\sigma(\varepsilon_t; \boldsymbol{\lambda})} \right) = \int f'_\sigma(\varepsilon_t; \boldsymbol{\lambda}) d\varepsilon_t = \sigma^{-1} f(x)|_{-\infty}^{\infty} = 0,$$

which follows by the definition of the density and assumption (A3) in Breidt et al. (1991). To simplify future computations, we begin by noting that

$$\mathbb{E}(z_s e_t) = \begin{cases} 0, & \text{if } s \neq t, \\ -1, & \text{if } s = t, \end{cases} \quad (19)$$

which follows from the assumptions on the density and strict exogeneity between all exogenous regressors and the error term. For $i = 1, \dots, r$, we can show that $\mathbb{E} \left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \phi_i} \right) = \mathbb{E}(-e_t v_{t-i}) = -\mathbb{E} \left(e_t \sum_{j=0}^{\infty} \alpha_j z_{t-i-j} \right) = 0$. Hence, we note that $\mathbf{V}_{t-1}$ and e_t are still independent as in Lanne and Saikkonen (2011). Consequently, we still find

$$\text{Cov} \left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \phi} \right) = \text{Cov}(-\mathbf{V}_{t-1} e_t) = \mathbb{E}(e_t^2) \mathbb{E}(\mathbf{V}_{t-1} \mathbf{V}_{t-1}') = \tilde{\mathcal{J}} \boldsymbol{\Gamma}_V,$$

where $\boldsymbol{\Gamma}_V$ denotes the autocovariance matrix of the vector $\mathbf{V}_{t-1}$. Because $\mathbf{V}_{t-1} e_t$ is uncorrelated, we have $\lim_{n \rightarrow \infty} \frac{1}{n} \text{Cov} \left(\sum_{t=r+1}^{T-s} \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \phi} \right) = \tilde{\mathcal{J}} \boldsymbol{\Gamma}_V$. Symmetrically, by using similar arguments, we can show for $i = 1, \dots, s$ that $\mathbb{E} \left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \varphi_i} \right) = \mathbb{E}(-e_t u_{t+i}) = -\mathbb{E} \left(e_t \sum_{j=0}^{\infty} \delta_j z_{t+i+j} \right) = 0$. That is, the independence of e_t and $\mathbf{U}_{t+1}$ is preserved through strict exogeneity. Letting $\boldsymbol{\Gamma}_U$ be the

autocovariance matrix of $\mathbf{U}_{t+1}$, we find that

$$\text{Cov}\left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\varphi}}\right) = \text{Cov}(-\mathbf{U}_{t+1}e_t) = \mathbb{E}(e_t^2)\mathbb{E}(\mathbf{U}_{t+1}\mathbf{U}_{t+1}') = \tilde{\mathcal{J}}\boldsymbol{\Gamma}_U.$$

Because $\mathbf{U}_{t+1}e_t$ is uncorrelated, we have $\lim_{n \rightarrow \infty} \frac{1}{n} \text{Cov}\left(\sum_{t=r+1}^{T-s} \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\varphi}}\right) = \tilde{\mathcal{J}}\boldsymbol{\Gamma}_U$. Lastly, we can apply the same logic for the parameter vector $\boldsymbol{\beta}$. Since for $i = 1, \dots, q$, we have that $\mathbb{E}\left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \beta_i}\right) = 0$ by the strict exogeneity of $x_{i,t}$ and ε_t . If we denote by $\boldsymbol{\Gamma}_X$, the autocovariance matrix of $\mathbf{X}_t$, it follows that

$$\text{Cov}\left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}}\right) = \text{Cov}(-\mathbf{X}_t e_t) = \mathbb{E}(e_t^2)\mathbb{E}(\mathbf{X}_t \mathbf{X}_t') = \tilde{\mathcal{J}}\boldsymbol{\Gamma}_X.$$

Because $\mathbf{X}_t e_t$ is uncorrelated, we have $\lim_{n \rightarrow \infty} \frac{1}{n} \text{Cov}\left(\sum_{t=r+1}^{T-s} \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}}\right) = \tilde{\mathcal{J}}\boldsymbol{\Gamma}_X$. We now characterize the covariances of the partials. To that end, we first notice that

$$\text{Cov}(z_{t-i}e_t, z_{k-j}e_k) = \begin{cases} \mathcal{I} + \tilde{\mathcal{J}} \sum_{m=1}^q \beta_m^2 \sigma_{x_m}^2 & \text{if } t = k, i = j = 0, \\ \mathcal{J} + \tilde{\mathcal{J}} \sum_{m=1}^q \beta_m^2 \sigma_{x_m}^2 & \text{if } t = k, i = j \neq 0, \\ 1 & \text{if } t \neq k, i = t - k, j = k - t, \\ 0 & \text{otherwise.} \end{cases} \quad (20)$$

Hence, using (19)-(20), we find that

$$\text{Cov}\left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \phi_i}, \frac{\partial g_k(\boldsymbol{\theta}_0)}{\partial \phi_j}\right) = \begin{cases} \gamma_V(i-j)\tilde{\mathcal{J}}, & \text{if } t = k, 1 \leq i \leq j \leq r, \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Cov}\left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \varphi_i}, \frac{\partial g_k(\boldsymbol{\theta}_0)}{\partial \varphi_j}\right) = \begin{cases} \gamma_U(i-j)\tilde{\mathcal{J}}, & \text{if } t = k, 1 \leq i \leq j \leq s, \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Cov} \left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \beta_i}, \frac{\partial g_k(\boldsymbol{\theta}_0)}{\partial \beta_j} \right) = \begin{cases} \sigma_{x_i}^2 \tilde{\mathcal{J}}, & \text{if } t = k, i = j \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Define $Q_m(i, j, a) \equiv \sum_{b=0}^{\infty} \delta_b \gamma_{x_m}(i + j + a + b)$. For the covariance matrix between $\partial g_t(\boldsymbol{\theta}_0)/\partial \phi$ and $\partial g_t(\boldsymbol{\theta}_0)/\partial \varphi$, first consider for $1 \leq i \leq r, 1 \leq j \leq s$:

$$\begin{aligned} \text{Cov} \left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \phi_i}, \frac{\partial g_k(\boldsymbol{\theta}_0)}{\partial \varphi_j} \right) &= \text{Cov} \left(\sum_{a=0}^{\infty} \alpha_a z_{t-i-a} e_t, \sum_{b=0}^{\infty} \delta_b z_{k+j+b} e_k \right) \\ &= \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \alpha_a \delta_b \text{Cov}(z_{t-i-a} e_t, z_{k+j+b} e_k) \\ &= \begin{cases} \alpha_{t-k-i} \delta_{t-k-j}, & \text{for } t > k, \\ \tilde{\mathcal{J}} \sum_{a=0}^{\infty} \alpha_a \sum_{m=1}^q \beta_m^2 Q_m(i, j, a) & \text{for } t = k, \\ 0 & \text{for } t < k. \end{cases} \end{aligned}$$

The element (i, j) of the matrix $\frac{1}{n} \text{Cov}(\partial L_T(\boldsymbol{\theta}_0)/\partial \phi, \partial L_T(\boldsymbol{\theta}_0)/\partial \varphi)$ is

$$\begin{aligned} &n \text{Cov} \left(\frac{1}{n} \sum_{t=r+1}^{T-s} \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \phi_i}, \frac{1}{n} \sum_{k=r+1}^{T-s} \frac{\partial g_k(\boldsymbol{\theta}_0)}{\partial \varphi_j} \right) \\ &= \frac{1}{n} \sum_{t=r+1}^{T-s} \sum_{k=r+1}^{T-s} \text{Cov} \left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \phi_i}, \frac{\partial g_k(\boldsymbol{\theta}_0)}{\partial \varphi_j} \right) \\ &= \frac{1}{n} \sum_{t=r+1}^{T-s} \sum_{k=r+1}^{T-s} \text{Cov}(-v_{t-i} e_t, -u_{k+j} e_k) \\ &= \frac{1}{n} \sum_{t=r+1}^{T-s} \sum_{k=r+1}^{T-s} \left(\mathbb{1}_{\{t>k\}} \alpha_{t-k-i} \delta_{t-k-j} + \mathbb{1}_{\{t=k\}} \tilde{\mathcal{J}} \sum_{a=0}^{\infty} \alpha_a \sum_{m=1}^q \beta_m^2 Q_m(i, j, a) \right) \\ &= \frac{1}{n} \left(\sum_{k=r+1}^{T-s-1} \sum_{t=0}^{T-s-k-i} \alpha_t \delta_{t+i-j} \right) + \tilde{\mathcal{J}} \sum_{a=0}^{\infty} \alpha_a \sum_{m=1}^q \beta_m^2 Q_m(i, j, a) \\ &\rightarrow \sum_{t=0}^{\infty} \alpha_t \delta_{t+i-j} + \tilde{\mathcal{J}} \sum_{a=0}^{\infty} \alpha_a \sum_{m=1}^q \beta_m^2 Q_m(i, j, a), \end{aligned}$$

where convergence of the first term follows from the geometric decay of the sequences $\{\alpha_t\}$ and $\{\delta_t\}$. Note that $\delta_{t+i-j} = 0$ for $t+i-j < 0$. The equalities follow from results presented earlier, the change of summands follows from imposing $t > k$.

Next, we consider the covariance between the partial derivatives of the log-likelihood with respect to the causal autoregressive parameters ϕ and the parameter vector of the exogenous variables β . That is,

$$\begin{aligned} \text{Cov}\left(\frac{\partial g_t(\theta_0)}{\partial \phi_i}, \frac{\partial g_k(\theta_0)}{\partial \beta_j}\right) &= \text{Cov}(-v_{t-i}e_t, -x_{j,k}e_k) \\ &= \mathbb{E}\left(\sum_{a=0}^{\infty} \alpha_a \varepsilon_{t-i-a} x_{j,k} e_k e_t\right) \\ &\quad + \mathbb{E}\left(\sum_{a=0}^{\infty} \alpha_a \sum_{m=1}^q \beta_m x_{m,t-i-a} x_{j,k} e_k e_t\right) \\ &= \begin{cases} \beta_j \tilde{\mathcal{J}} \sum_{a=0}^{\infty} \alpha_a \gamma_{x_j}(i+a) & \text{for } t = k, 1 \leq i \leq r, \\ 0 & \text{for } t \neq k, 1 \leq i \leq r. \end{cases} \end{aligned}$$

Note that this outcome is independent of time t . Symmetrically, we can compute the covariance between the partial derivatives of the log-likelihood with respect to the noncausal autoregressive parameters φ and the parameter vector of the exogenous variables β :

$$\begin{aligned} \text{Cov}\left(\frac{\partial g_t(\theta_0)}{\partial \varphi_i}, \frac{\partial g_k(\theta_0)}{\partial \beta_j}\right) &= \text{Cov}(-u_{t+i}e_t, -x_{j,k}e_k) \\ &= \mathbb{E}\left(\sum_{b=0}^{\infty} \delta_b \varepsilon_{t+i+b} x_{j,k} e_k e_t\right) \\ &\quad + \mathbb{E}\left(\sum_{b=0}^{\infty} \delta_b \sum_{m=1}^q \beta_m x_{m,t+i+b} x_{j,k} e_k e_t\right) \\ &= \begin{cases} \beta_j \tilde{\mathcal{J}} \sum_{b=0}^{\infty} \delta_b \gamma_{x_j}(i+b) & \text{for } t = k, 1 \leq i \leq s, \\ 0 & \text{for } t \neq k, 1 \leq i \leq s. \end{cases} \end{aligned}$$

The proof of asymptotic normality is similar to Breidt et al. (1991) and Lanne and Saikkonen (2011). Define $\mathbf{M} = \text{diag}(\mathbf{\Sigma}, \mathbf{\Omega})$, $\mathbf{W}_t = \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ and note that $n \equiv (T - p)$. By the Cramér-Wold theorem, it suffices to show that for any vector $\mathbf{a}$ of appropriate size,

$$\frac{1}{\sqrt{n}} \sum_{t=r+1}^{T-s} \mathbf{a}' \mathbf{W}_t \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{a}' \mathbf{M} \mathbf{a}). \quad (21)$$

Define the sequence of $(p + q + d + 1)$ dimensional random vectors $\{\mathbf{W}_{tm}, t \in \mathbb{Z}\}$ to be the partials defined in Appendix A, where v_t , u_t and all $x_{i,t}$ for $i = 1, \dots, q$ are replaced by their representation $u_t = \sum_{j=0}^{\infty} \delta_j z_{t+j}$ and $v_t = \sum_{j=0}^{\infty} \alpha_j z_{t-j}$ and assumption (II) with the sums truncated at a large positive integer m , i.e., $v_t^{(m)} = \sum_{j=0}^m \alpha_j z_{t-j}$, $u_t^{(m)} = \sum_{j=0}^m \delta_j z_{t+j}$ and $x_{i,t}^{(m)} = c_i + \sum_{j=-m}^m \rho_{i,j} \eta_{i,t-j}$.

It can be verified that $\mathbb{E}(\mathbf{W}_t) = \mathbf{0}$ and $\gamma_{\mathbf{W}_t}(0) + 2 \sum_{j=1}^{\infty} \gamma_{\mathbf{W}_t}(j) \neq \mathbf{0}$. This result also holds for $\mathbf{W}_t$ replaced by $\mathbf{W}_{tm}$. Let $\mathbf{M}_m$ be the matrix corresponding to $\mathbf{M}$, obtained by truncating u_t , v_t and $\mathbf{X}_t$. Then the stationary sequence $\{\mathbf{W}_{tm}, t \in \mathbb{Z}\}$ is $\max\{m + p, 2m\}$ dependent.²⁰ Now that we verified the conditions, we can apply Theorem 6.4.2 in Brockwell and Davis (1991) to obtain

$$\frac{1}{\sqrt{n}} \sum_{t=r+1}^{T-s} \mathbf{a}' \mathbf{W}_{tm} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{a}' \mathbf{M}_m \mathbf{a}).$$

Now, it follows that for $m \rightarrow \infty$, $\mathbf{W}_{tm} \rightarrow \mathbf{W}_t$ (by definition) and thus $\mathbf{M}_m \rightarrow \mathbf{M}$. Because

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \text{Var} \left(\frac{1}{\sqrt{n}} \sum_{t=r+1}^{T-s} (\mathbf{a}' \mathbf{W}_{tm} - \mathbf{a}' \mathbf{W}_t) \right) = 0,$$

the convergence in (21) is immediate from Proposition 6.3.9 in Brockwell and Davis (1991). The positive definiteness of $\mathbf{\Sigma}$ can be established similar to the proof in Breidt et al. (1991). In the

²⁰The $m + p$ follows from writing u_t and v_t in their truncated representation, $2m$ follows from the processes in $\mathbf{X}_t$ which have a two-sided MA representation truncated by m at both sides.

MARX case, the block matrix $\mathbf{\Sigma}$ is given as

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_{13} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} & \mathbf{\Sigma}_{23} \\ \mathbf{\Sigma}_{31} & \mathbf{\Sigma}_{32} & \mathbf{\Sigma}_{33} \end{bmatrix}.$$

In a first step, let us focus on the submatrix $\tilde{\mathbf{\Sigma}}_1$ given by

$$\tilde{\mathbf{\Sigma}}_1 = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \text{ and partition it as } \tilde{\mathbf{\Sigma}}_1 = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{bmatrix},$$

where $\mathbf{A}$ is $r \times r$, $\mathbf{C}$ is $s \times s$ and $\mathbf{B}$ is $r \times s$. Consider the vectors $\mathbf{P} = [P_1, \dots, P_r]'$ and $\mathbf{S} = [S_1, \dots, S_s]'$ defined by $\mathbf{P}_t = \sum_{a=0}^{\infty} \alpha_a z_{t-a} e_0$ for $t = 1, \dots, r$ and $\mathbf{S}_t = \sum_{b=0}^{\infty} \delta_b z_{t-b} e_0$ for $t = 1, \dots, s$. It can easily be verified that the covariance matrices of $\mathbf{P}$ and $\mathbf{S}$, denoted $\mathbf{\Sigma}_{PP}$ and $\mathbf{\Sigma}_{SS}$, are equal to $\mathbf{A}$ and $\mathbf{C}$. From (20), it follows that

$$\begin{aligned} \text{Cov}(P_i, S_j) &= \text{Cov} \left(\sum_{a=0}^{\infty} \alpha_a z_{i-a} e_0, \sum_{b=0}^{\infty} \delta_b z_{j-b} e_0 \right) \\ &= \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} \alpha_a \delta_b \mathbb{E}(z_{i-a} z_{j-b} e_0^2) \\ &= \sum_{a=0}^{\infty} \alpha_a \delta_{b+i-j} (\mathcal{J} + \tilde{\mathcal{J}} \sum_{m=1}^q \beta_m^2 \sigma_m^2). \end{aligned}$$

We know that $\mathcal{J} > 1$ by condition (A5) of Andrews et al. (2006). We have that $\tilde{\mathcal{J}} \sum_{m=1}^q \beta_m^2 \sigma_m^2 = \mathcal{J} \left(\frac{\sum_{m=1}^q \beta_m^2 \sigma_m^2}{\sigma^2} \right) > 0$, which in turn implies that $(\mathcal{J} + \tilde{\mathcal{J}} \sum_{m=1}^q \beta_m^2 \sigma_m^2) > 1$. Similar to Breidt et al. (1991), we exploit that the matrices $\mathbf{A}$ and $\mathbf{C}$ are positive definite since there is no linear dependence within the vectors $\mathbf{P}$ and $\mathbf{S}$. We proceed by proving the positive definiteness of $\tilde{\mathbf{\Sigma}}$ by showing that the Schur Complement of the block $\mathbf{A}$ of the matrix $\tilde{\mathbf{\Sigma}}$ given as $\mathbf{C} - \mathbf{B}' \mathbf{A}^{-1} \mathbf{B}$ is positive definite. We know that the covariance matrix of $\mathbf{S} - \mathbf{\Sigma}_{SP} \mathbf{\Sigma}_{PP}^{-1}$, i.e. $\mathbf{C} - (\mathcal{J} +$

$\tilde{\mathcal{J}} \sum_{m=1}^q \beta_m^2 \sigma_m^2) \mathbf{B}' \mathbf{A}^{-1} \mathbf{B}$, is positive semidefinite and hence for a nonzero vector $\mathbf{c} \in \mathbb{R}^s$ with $\mathbf{B}\mathbf{c} \neq \mathbf{0}$, we have that

$$\mathbf{c}'(\mathbf{C} - \mathbf{B}' \mathbf{A}^{-1} \mathbf{B}) \mathbf{c} > \mathbf{c}'(\mathbf{C} - (\mathcal{J} + \tilde{\mathcal{J}} \sum_{m=1}^q \beta_m^2 \sigma_m^2) \mathbf{B}' \mathbf{A}^{-1} \mathbf{B}) \mathbf{c} \geq 0.$$

Alternatively, if $\mathbf{B}\mathbf{c} = \mathbf{0}$, we have that

$$\mathbf{c}'(\mathbf{C} - \mathbf{B}' \mathbf{A}^{-1} \mathbf{B}) \mathbf{c} = \mathbf{c}' \mathbf{C} \mathbf{c} > 0,$$

by the positive definiteness of $\mathbf{C}$. Hence, now that we established positive definiteness of $\tilde{\Sigma}_1$, we can repartition the matrix Σ as

$$\Sigma = \begin{bmatrix} \tilde{\Sigma}_1 & \tilde{\Sigma}_2 \\ \tilde{\Sigma}_2' & \tilde{\Sigma}_3 \end{bmatrix},$$

where $\tilde{\Sigma}_1$ is $(r+s) \times (r+s)$, $\tilde{\Sigma}_2 = [\Sigma_{12}, \Sigma_{23}]'$ is $(r+s) \times q$ and $\tilde{\Sigma}_3 = \Sigma_{33}$ is $q \times q$. Since $\Sigma_{33} = \text{diag}(\sigma_1^2, \dots, \sigma_m^2)$, we have that for a nonzero vector $\mathbf{c} \in \mathbb{R}^q$, $\mathbf{c}' \Sigma_{33} \mathbf{c} = c_1^2 \sigma_1^2 + \dots + c_m^2 \sigma_m^2 > 0$. Hence, as we know that $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_3$ are positive definite, it is sufficient to show that the Schur complement of the block $\tilde{\Sigma}_1$ of the matrix Σ is positive definite, which can be established analogous to the case above. The positive definiteness of Ω follows from condition (A6) in Andrews et al. (2006).

A.2 Proof of Theorem 1

We first present the second partial derivatives of the function $g_t(\theta)$. Set $h(x; \lambda) = f'(x; \lambda)/f(x; \lambda)$, such that

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

which can easily be verified using the quotient rule. Let $\mathbf{Y}_t$ be the $(r \times s)$ matrix with elements y_{t-i+j} . Write $\tilde{v}_t = v_t(\boldsymbol{\varphi})$ and $\tilde{u}_t = u_t(\boldsymbol{\phi})$ and thus $\tilde{\mathbf{V}}_{t-1} = [\tilde{v}_{t-1}, \dots, \tilde{v}_{t-r}]'$ and $\tilde{\mathbf{U}}_{t+1} = [\tilde{u}_{t+1}, \dots, \tilde{u}_{t+s}]'$ to simplify notation. Similarly, $\tilde{\varepsilon}_t = \tilde{v}_t - \phi_1 \tilde{v}_{t-1} - \dots - \phi_r \tilde{v}_{t-r} = \tilde{u}_t - \varphi_1 \tilde{u}_{t+1} - \dots - \varphi_s \tilde{u}_{t+s}$ denotes ε_t evaluated at an arbitrary point in the permissible parameter space, not the true one. Then, the second partial derivatives in the MARX case can be obtained through differentiation, similar to Lanne and Saikkonen (2011) and Breidt et al. (1991):

$$\begin{aligned}
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\phi} \partial \boldsymbol{\phi}' &= \sigma^{-2} h'(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\mathbf{V}}_{t-1} \tilde{\mathbf{V}}_{t-1}', \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}' &= \sigma^{-2} h'(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\mathbf{U}}_{t+1} \tilde{\mathbf{U}}_{t+1}', \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}' &= \sigma^{-2} h'(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \mathbf{X}_t \mathbf{X}_t', \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \sigma^2 &= 2\sigma^{-3} h(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\varepsilon}_t + \sigma^{-4} h'(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\varepsilon}_t^2 + \sigma^{-2}, \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}' &= \frac{1}{f(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda})} \frac{\partial^2 f(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} \\
&\quad - \frac{1}{f^2(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda})} \left(\frac{\partial f(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \right) \left(\frac{\partial f(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \right)', \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\phi} \partial \boldsymbol{\varphi}' &= \sigma^{-2} h'(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\mathbf{V}}_{t-1} \tilde{\mathbf{U}}_{t+1}' + \sigma^{-1} h(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \mathbf{Y}_t, \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\phi} \partial \boldsymbol{\beta}' &= \sigma^{-2} h'(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\mathbf{V}}_{t-1} \mathbf{X}_t', \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\phi} \partial \sigma &= \sigma^{-3} h'(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\varepsilon}_t \tilde{\mathbf{V}}_{t-1} + \sigma^{-2} h(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\mathbf{V}}_{t-1}, \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\phi} \partial \boldsymbol{\lambda}' &= -\sigma^{-1} \tilde{\mathbf{V}}_{t-1} \partial h(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) / \partial \boldsymbol{\lambda}', \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\varphi} \partial \boldsymbol{\beta}' &= \sigma^{-2} h'(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\mathbf{U}}_{t+1} \mathbf{X}_t', \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\varphi} \partial \sigma &= \sigma^{-3} h'(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\varepsilon}_t \tilde{\mathbf{U}}_{t+1} + \sigma^{-2} h(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\mathbf{U}}_{t+1}, \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\varphi} \partial \boldsymbol{\lambda}' &= -\sigma^{-1} \tilde{\mathbf{U}}_{t+1} \partial h(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) / \partial \boldsymbol{\lambda}', \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\beta} \partial \sigma &= \sigma^{-3} h'(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \tilde{\varepsilon}_t \mathbf{X}_t + \sigma^{-2} h(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) \mathbf{X}_t, \\
\partial^2 g_t(\boldsymbol{\theta}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\lambda}' &= -\sigma^{-1} \mathbf{X}_t \partial h(\sigma^{-1} \tilde{\varepsilon}_t; \boldsymbol{\lambda}) / \partial \boldsymbol{\lambda}',
\end{aligned}$$

$$\partial^2 g_t(\boldsymbol{\theta})/\partial\sigma\partial\boldsymbol{\lambda}' = -\sigma^{-2}\tilde{\varepsilon}_t\partial h(\sigma^{-1}\tilde{\varepsilon}_t; \boldsymbol{\lambda})/\partial\boldsymbol{\lambda}'.$$

It can be verified that $\mathbb{E}(\partial^2 g_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}') = -\text{diag}(\boldsymbol{\Sigma}, \boldsymbol{\Omega})$. The proof for consistency is exactly the same as in Lanne and Saikkonen (2011). That is, similar to Andrews et al. (2006), we use the Taylor expansion

$$\begin{aligned} \sum_{t=r+1}^{T-s} \left[g_t(\boldsymbol{\theta}_0 + T^{-1/2}\mathbf{c}) - g_t(\boldsymbol{\theta}_0) \right] &= \frac{1}{\sqrt{T}} \sum_{t=r+1}^{T-s} \mathbf{c}' \frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} + \frac{1}{2T} \sum_{t=r+1}^{T-s} \mathbf{c}' \frac{\partial^2 g_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \mathbf{c} \\ &\quad + \frac{1}{2T} \sum_{t=r+1}^{T-s} \mathbf{c}' \left(\frac{\partial^2 g_t(\boldsymbol{\theta}_T^*(\mathbf{c}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial^2 g_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \right) \mathbf{c}, \end{aligned}$$

where $\mathbf{c} \in \mathbb{R}^{p+q+1+d}$ and the argument $\boldsymbol{\theta}_T^*(\mathbf{c})$ in the matrix of second partial derivatives means that each row is evaluated at an intermediate point lying between the true parameter value $\boldsymbol{\theta}_0$ and $T^{-1/2}\mathbf{c}$. If $\|\cdot\|$ denotes the Euclidian norm we have $\sup_{\mathbf{c} \in \mathbf{C}} \|\boldsymbol{\theta}_T^*(\mathbf{c}) - \boldsymbol{\theta}_0\| \rightarrow 0$ for any compact set $\mathbf{C} \subset \mathbb{R}^{p+q+1+d}$. Using the dominance conditions (A7) in Davis et al. (1992), arguments similar to Breidt et al. (1991, p. 186-190) and assumption (I) in this paper, it can be shown that a uniform law of large numbers for stationary ergodic processes applies to $\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ over any small enough compact neighborhood $\boldsymbol{\theta}_0$. We can conclude that

$$\frac{1}{T} \sum_{t=r+1}^{T-s} \mathbf{c}' \left(\frac{\partial^2 g_t(\boldsymbol{\theta}_T^*(\mathbf{c}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial^2 g_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \right) \mathbf{c} \xrightarrow{p} 0,$$

for $\mathbf{c} \in \mathbf{C}$. As in the proof of Theorem 1 of Andrews et al. (2006), we can make use of Remark 1 of Davis et al. (1992) and complete the proof.

A.3 Approximate Likelihood Function

Define $\mathbf{b} = \beta'\tilde{\mathbf{x}}$ such that $\mathbf{z} = \mathbf{B}\mathbf{A}\mathbf{y} - \beta'\tilde{\mathbf{x}} = \mathbf{B}\mathbf{A}\mathbf{y} - \mathbf{b}$. Assume $\mathbf{B}$ and $\mathbf{A}$ are invertible. We are interested in the inverse transformation, i.e. $\mathbf{y} = \mathbf{Q}(\mathbf{z} + \mathbf{b})$, where $\mathbf{Q} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. Let $\mathbf{Q}$ be

a (2×2) matrix, then we have

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} q_1 & q_2 \\ q_3 & q_4 \end{bmatrix} \left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \right), \quad (22)$$

with the following functions $y_1 = g_1(z_1, z_2) = q_1 z_1 + q_2 z_2 + b_1$ and $y_2 = g_2(z_1, z_2) = q_3 z_1 + q_4 z_2 + b_2$.

The Jacobian is given as the matrix of all partial derivatives from y to z , i.e.

$$\mathbf{J} = \begin{bmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} \\ \frac{\partial y_2}{\partial z_1} & \frac{\partial y_2}{\partial z_2} \end{bmatrix} = \begin{bmatrix} q_1 & q_2 \\ q_3 & q_4 \end{bmatrix} = \mathbf{Q}. \quad (23)$$

Then the joint density of y_1 and y_2 is given as $f_{y_1, y_2}(y_1, y_2) = \frac{1}{|\det(\mathbf{Q})|} f_{z_1, z_2}(\mathbf{B}\mathbf{A}\mathbf{y} - \mathbf{b})$. This result can be generalized to higher orders (see e.g., Casella and Berger, 2002, p. 185). From Proposition 1 we know that the information sets (i) and (vi) are observationally equivalent. Using this result and $\mathbf{Q} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, we find that the probability density of the process y_t can be represented in the following way:

$$\begin{aligned} f_{y; \lambda}(\mathbf{y}) &= \frac{1}{|\det(\mathbf{Q})|} f_z(\mathbf{B}\mathbf{A}\mathbf{y} - \mathbf{b}; \lambda) \\ &= |\det(\mathbf{A})| |\det(\mathbf{B})| h_V(\mathbf{B}\mathbf{A}\mathbf{y} - \mathbf{b}) f_\varepsilon(\mathbf{B}\mathbf{A}\mathbf{y} - \mathbf{b}; \lambda) h_U(\mathbf{B}\mathbf{A}\mathbf{y} - \mathbf{b}) \\ &= |\det(\mathbf{A})| h_V(\mathbf{B}\mathbf{A}\mathbf{y} - \mathbf{b}) \left(\prod_{t=r+1}^{T-s} f_\sigma(\mathbf{B}\mathbf{A}\mathbf{y} - \mathbf{b}; \lambda) \right) h_U(\mathbf{B}\mathbf{A}\mathbf{y} - \mathbf{b}) \\ &= h_V(\varphi(L^{-1})y_1, \dots, \varphi(L^{-1})y_r) \left(\prod_{t=r+1}^{T-s} f_\sigma(\phi(L)\varphi(L^{-1})y_t - \beta' \mathbf{X}_t; \lambda) \right) \\ &\quad h_U(\phi(L)y_{T-s+1}, \dots, \phi(L)y_T) |\det(\mathbf{A})|, \end{aligned} \quad (24)$$

where $\mathbf{A}$ and $\mathbf{B}$ are two nonsingular matrices with $\det(\mathbf{B}) = 1$; h_V and h_U are the joint densities of $(v_1, \dots, v_r)$ and $(u_{T-s+1}, \dots, u_T)$ respectively. Independence of the blocks $(v_1, \dots, v_r)$, $(\varepsilon_{r+1}, \dots, \varepsilon_{T-s})$ and $(u_{T-s+1}, \dots, u_T)$ is applied in the second equality and the definition of the

filtered values as presented in (4) and (5) in the fourth equality. Since $\det(\mathbf{A})$ is independent of sample size, the density of y_t can be approximated by the second term in (24).

A.4 From Transfer Function Model to MARX

For expository purposes, we consider a single explanatory variable denoted x_t^* . The transfer function model is given by

$$y_t = \psi^*(L)x_t^* + n_t, \quad (25)$$

where n_t is a noise process assumed to follow a stationary AR process, $a(L)n_t = \varepsilon_t^*$. The ARX (or ARDL) model can be motivated from (25) by assuming that the transfer function operator can be expressed in a rational factorization as $\psi^*(z) = a(z)^{-1}\theta^*(z)$. Multiplying (25) by $a(L)$ yields

$$\begin{aligned} a(L)y_t &= \theta^*(L)x_t^* + a(L)n_t \\ &= \theta^*(L)x_t^* + \varepsilon_t^*, \end{aligned} \quad (26)$$

which is the usual $\text{ARX}(p, k)$ model representation when $\deg(a(z)) = p$ and $\deg(\theta^*(z)) = k$. If all roots of $a(z)$ lie outside the unit circle, the process is stationary which implies that estimation and inference can directly be conducted. Breidt et al. (1991) consider the more complex case in which r roots lie outside the unit circle and s inside ($r + s = p$) and propose to factorize the polynomial to obtain

$$\phi(L)\varphi^*(L)y_t = \theta^*(L)x_t^* + \varepsilon_t^*.$$

Lanne and Saikkonen (2011) propose to rewrite $\varphi^*(z)$ in terms of the lead operator and obtain the relation $\varphi^*(z) = -\varphi_s^* z^s \varphi(z^{-1})$. By rearranging terms, we find

$$\phi(L)\varphi(L^{-1})y_t = \left(-\frac{1}{\varphi_s^*} + \frac{\theta_1^*}{\varphi_s^*}L + \dots + \frac{\theta_k^*}{\varphi_s^*}L^k \right) x_{t+s}^* - \frac{1}{\varphi_s^*}\varepsilon_{t+s}^*$$

$$= \theta(L)x_t + \varepsilon_t. \quad (27)$$

In case only a contemporaneous value of x_t enters the system, take $\psi^*(z) = a(z)^{-1}\beta$. Note that the derivation can easily be extended to q regressors by defining $\psi^*(z) = [\psi_1^*(z), \dots, \psi_q^*(z)]'$ and considering X_t^* . In the distributed lag case take $\psi^*(z) = a(z)^{-1}\theta^*(z)$ with $\theta^*(z) = [\theta_1^*(z), \dots, \theta_k^*(z)]'$; in the contemporaneous case define $\psi^*(z) = a(z)^{-1}\beta$ with $\beta \in \mathbb{R}^q$. In case one wants to allow for (mixed) dynamics in the exogenous regressors, it seems more natural to model such a process as a VAR. The mixed VAR model (see e.g., Lanne and Saikkonen, 2013; Davis and Song, 2012) accommodates this structure.

B Auxiliary Proposition with Proof

Proposition 1. *For an MARX(r, s, q) model, the following information sets are equivalent:*

- (i) $(y_1, \dots, y_T, \mathbf{X}_{r+1}, \dots, \mathbf{X}_{T-s})$
- (ii) $(y_1, \dots, y_r, u_{r+1}, \dots, u_T, \mathbf{X}_{r+1}, \dots, \mathbf{X}_{T-s})$
- (iii) $(v_1, \dots, v_{T-s}, y_{T-s+1}, \dots, y_T, \mathbf{X}_{r+1}, \dots, \mathbf{X}_{T-s})$
- (iv) $(y_1, \dots, y_r, \varepsilon_{r+1}, \dots, \varepsilon_{T-s}, u_{T-s+1}, \dots, u_T)$
- (v) $(v_1, \dots, v_r, \varepsilon_{r+1}, \dots, \varepsilon_{T-s}, y_{T-s+1}, \dots, y_T)$
- (vi) $(v_1, \dots, v_r, \varepsilon_{r+1}, \dots, \varepsilon_{T-s}, u_{T-s+1}, \dots, u_T)$

Additionally, the following information sets are equivalent:

- (i') $(y_1, \dots, y_T)$
- (ii') $(y_1, \dots, y_r, u_{r+1}, \dots, u_T)$
- (iii') $(v_1, \dots, v_{T-s}, y_{T-s+1}, \dots, y_T)$

Proof of Proposition 1

Let $\sim$ denote equivalence in information sets. To show that $(i), (ii)$ and (iii) are equivalent is similar to showing that $(i'), (ii')$ and (iii') are equivalent. We prove $(i') \sim (ii'), (i') \sim (iii'), (ii) \sim (iv), (iii) \sim (v)$ and $(i) \sim (vi)$.

Case 1: $(i') \sim (ii')$. Using (4), $(ii') (y_1, \dots, y_r, u_{r+1}, \dots, u_T) = (y_1, \dots, y_r, \phi(L)y_{r+1}, \dots, \phi(L)y_T)$. Since $u_{r+1} = y_{r+1} - \phi_1 y_r - \dots - \phi_r y_1$ with $y_1, \dots, y_r$ and u_{r+1} known, y_{r+1} is known. The same reasoning can be recursively applied to u_{r+2} up to u_T , leading to the desired result.

Case 2: $(i') \sim (iii')$. By (5), $(iii') (v_1, \dots, v_{T-s}, y_{T-s+1}, \dots, y_T) = (\varphi(L^{-1})y_1, \dots, \varphi(L^{-1})y_{T-s}, y_{T-s+1}, \dots, y_T)$. Since $v_{T-s} = y_{T-s} - \varphi_1 y_{T-s+1} - \dots - \varphi_s y_T$ with $y_{T-s+1}, \dots, y_T$ and v_{T-s} known, y_{T-s} is known. The same reasoning can be recursively applied to v_{T-s-1} up to v_1 , leading to the desired result. Hence, since $(i'), (ii')$ and (iii') are equivalent, we know that $(i), (ii)$ and (iii) are as well (as all information sets are augmented with the same information).

Case 3: $(ii) \sim (iv)$. We have $(iv) (y_1, \dots, y_r, \varepsilon_{r+1}, \dots, \varepsilon_{T-s}, u_{T-s+1}, \dots, u_T) = (y_1, \dots, y_r, \varphi(L^{-1})u_{r+1} - \beta' \mathbf{X}_{r+1}, \dots, \varphi(L^{-1})u_{T-s} - \beta' \mathbf{X}_{T-s}, u_{T-s+1}, \dots, u_T)$ by using the second equality in equation (4). Since $\varepsilon_{T-s} = u_{T-s} - \varphi_1 u_{T-s+1} - \dots - \varphi_s u_T - \beta' \mathbf{X}_{T-s}$ with $u_{T-s+1}, \dots, u_T$, $\mathbf{X}_{T-s}$ and ε_{T-s} known, u_{T-s} is known. The same reasoning can be recursively applied to u_{T-s-1} up to u_{r+1} .

Case 4: $(iii) \sim (v)$. We have $(v) (v_1, \dots, v_r, \varepsilon_{r+1}, \dots, \varepsilon_{T-s}, y_{T-s+1}, \dots, y_T) = (v_1, \dots, v_r, \phi(L)v_{r+1} - \beta' \mathbf{X}_{r+1}, \dots, \phi(L)v_{T-s} - \beta' \mathbf{X}_{T-s}, y_{T-s+1}, \dots, y_T)$ by using the second equality in equation (5). Since $\varepsilon_{r+1} = v_{r+1} - \phi_1 v_r - \dots - \phi_r v_1 - \beta' \mathbf{X}_{r+1}$ with $v_1, \dots, v_r$, $\mathbf{X}_{r+1}$ and ε_{r+1} known, v_{r+1} is known. This reasoning can be recursively applied to v_{r+2} up to v_{T-s} .

Case 5: $(i) \sim (vi)$. To show: $(y_1, \dots, y_T, \mathbf{X}_r, \dots, \mathbf{X}_{T-s}) \sim (v_1, \dots, v_r, \varepsilon_{r+1}, \dots, \varepsilon_{T-s}, u_{T-s+1}, \dots, u_T)$. We denote the vector corresponding to the information set (i) by $\tilde{\mathbf{y}}$ and (ii) by $\mathbf{z}$ and use results similar to Lanne and Saikkonen (2011). Define $\mathbf{w} = [v_1, \dots, v_{T-s}, u_{T-s+1}, \dots, u_T]'$ and

$\mathbf{y} = [y_1, \dots, y_T]'$. Then

$$\begin{bmatrix} v_1 \\ \vdots \\ v_{T-s} \\ u_{T-s+1} \\ \vdots \\ v_T \end{bmatrix} = \begin{bmatrix} y_1 - \varphi_1 y_2 - \dots - \varphi_s y_{s+1} \\ \vdots \\ y_{T-s} - \varphi_1 y_{T-s+1} - \dots - \varphi_s y_T \\ y_{T-s+1} - \phi_1 y_{T-s} - \dots - \phi_r y_{T-s+1-r} \\ \vdots \\ y_T - \phi_1 y_{T-1} - \dots - \phi_r y_{T-r} \end{bmatrix} = \mathbf{A} \begin{bmatrix} y_1 \\ \vdots \\ y_{T-s} \\ y_{T-s+1} \\ \vdots \\ y_T \end{bmatrix},$$

which can be written as $\mathbf{w} = \mathbf{A}\mathbf{y}$. Now define $\tilde{\mathbf{x}} = [\underbrace{0, \dots, 0}_{r \text{ times}}, \mathbf{X}_{r+1}, \dots, \mathbf{X}_{T-s}, \underbrace{0, \dots, 0}_{s \text{ times}}]'$. Similarly, we form the following system of equations for $\mathbf{z}$, since (6) shows that $\mathbf{z} = \mathbf{B}\mathbf{w} - \beta'\tilde{\mathbf{x}}$. Combining both systems of equations, we find that $\mathbf{z}$ and $\tilde{\mathbf{y}}$ are related as $\mathbf{z} = \mathbf{B}\mathbf{A}\mathbf{y} - \beta'\tilde{\mathbf{x}}$, where $\mathbf{y}$ and $\tilde{\mathbf{x}}$ combined form the information set $\tilde{\mathbf{y}}$. Since $\mathbf{B}$, $\mathbf{A}$ and β only contain known parameters, this shows that (i) and (ii) are equivalent. All cases combined show that information sets (i) – (vi) are equivalent.

$$\begin{bmatrix} v_1 \\ \vdots \\ v_r \\ v_{r+1} - \phi_1 v_r - \dots - \phi_r v_1 - \beta' \mathbf{X}_{r+1} \\ \vdots \\ v_{T-s} - \phi_1 v_{T-s-1} - \dots - \phi_r v_{T-s-r} - \beta' \mathbf{X}_{T-s} \\ u_{T-s+1} \\ \vdots \\ u_T \end{bmatrix} = \mathbf{B} \begin{bmatrix} v_1 \\ \vdots \\ v_r \\ v_{r+1} \\ \vdots \\ v_{T-s} \\ u_{T-s+1} \\ \vdots \\ u_T \end{bmatrix} - \beta' \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{X}_{r+1} \\ \vdots \\ \mathbf{X}_{T-s} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

References

- Andrews, B., Breidt, F. & Davis, R. (2006). Maximum Likelihood Estimation For All-Pass Time Series Models. *Journal of Multivariate Analysis*, 97, 1638-1659.
- Breidt, F., Davis, R., Lii, K. & Rosenblatt, M. (1991). Maximum Likelihood Estimation for Noncausal Autoregressive Processes. *Journal of Multivariate Analysis*, 36, 175-198.
- Brockwell, P. & Davis, R. (1991). *Time Series: Theory and Methods*. Springer-Verlag New York, Second Edition.
- Casella, G. & Berger, R. (2002). *Statistical Inference*. Thomson Learning, Second Edition.
- Davis, R., Knight, K. & Liu, J. (1992). M-Estimation for Autoregressions with Infinite Variance. *Stochastic Processes and Their Applications*, 40, 145-180.
- Davis, R. & Song, L. (2012). Noncausal Vector AR Processes with Application to Economic Time Series. Discussion Paper Colombia University.
- Lanne, M. & Saikkonen, P. (2011). Noncausal Autoregressions for Economic Time Series. *Journal of Time Series Econometrics*, 3(3), 1-32.
- Lanne, M. & Saikkonen, P. (2013). Noncausal Vector Autoregression. *Econometric Theory*, 29(3), 447-481.