Mixed Causal-Noncausal Autoregressions with Strictly Exogenous Regressors

João Victor Issler* **Alain Hecq†** **Sean Telg‡**

November 22, 2016

Abstract

Some authors propose mixed autoregressive causal-noncausal (MAR) models to estimate economic relationships involving expectations variables. These structural equations usually imply explosive roots in their autoregressive part but have stationary forward solutions. In previous work, possible exogenous variables in economic relationships are substituted into the error term and are assumed to also follow an MAR process to ensure the MAR structure of the variable of interest. To allow for the impact of exogenous variables directly, we instead consider a MARX representation which allows for the inclusion of strictly exogenous regressors. We develop the asymptotic distribution of the MARX parameters. 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 investigate the link between commodity prices and the US exchange rate.

Keywords: Mixed causal-concausal process, non-Gaussian errors, identification, rational expectations models, commodity prices.

JEL codes: C22, E31, E37.

*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.

†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.

‡Corresponding author: Sean Telg, Maastricht University, School of Business and Economics, Department of Quantitative Economics, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: j.telg@maastrichtuniversity.nl.

1 Introduction

The usefulness of mixed causal-noncausal autoregressive (MAR) models in time series econometrics can be explained to three reasons. First, Gouriéroux and Zakoïan (2016) as well as Hencic and Gouriéroux (2014) and Hecq, Lieb and Telg (2016a) demonstrate how the inclusion of noncausal autoregressive terms can generate dynamic patterns like speculative bubbles and asymmetric cycles that beforehand could only be generated using complex nonlinear models. Second, several authors (see e.g., Lanne, Luoto and Saikkonen, 2012; Lanne, Nyberg and Saarinen, 2012) have shown that allowing for noncausality might improve forecast performances. Third, the MAR representation of an economic variable can be interpreted as a solution of a rational expectation model (Lanne and Saikkonen, 2011). Whereas the conventional causal autoregressive model takes only the fundamental solution of a system into account, the mixed causal-noncausal autoregressive (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, Barigozzi and Capasso, 2011) and explicitly consider that the information available to economic agents is larger than the one of econometricians.

In all three cases, it is justified to believe that the inclusion of exogenous terms in the model might further improve the relevance of the MAR model for modelling and forecasting economic processes. As an example, Lof and Nyberg (2015) estimate MAR models for commodity prices and show that adding in a second step the exchange rate leads to improved forecasts. More generally, a certain peak or trough in a series might not be due to the dynamic structure of the series itself, but because of a shock in another variable or in the economy as a whole. This is what we notice when estimating an ARMA model for macroeconomic time series without being able to capture the 2008 financial crisis drop. The inclusion of exogenous variables with the same dramatic feature can improve the forecast performance of a causal autoregressive model (see e.g., Corradi and Swanson, 2006), a result which is likely to extend to the mixed case as well. In a rational expectation framework, economic agents exploit all information available in the market, hence also information on exogenous forcing variables. An example is the new Keynesian Phillips curve (NKPC) where the solution of the model for the inflation rate depends on the output gap (see e.g. Pesaran, 2015, p. 475).

For the above reasons, this paper studies the MAR model enriched with strictly exogenous variables (MARX). We show how the MARX parameters can be estimated by the method of maximum likelihood (ML) and we provide closed form solutions of the corresponding standard errors in the Student's t-likelihood framework. We show that MARX models have the appealing feature to be (potentially) identifiable under Gaussianity, while their MAR counterparts are not. By means of Monte Carlo simulations, we evaluate the performance of the ML estimator and a proposed model selection for the MARX model based on information criteria. We investigate the relationship between commodity prices indices representative of the global market ($cp_t = \ln CP_t$) and the US nominal exchange rate ($s_t = \ln S_t$).[1] Our starting point is a rational expectation (RE) model with forward and backward components similar to the NKPC, such as the simple one-lag one-lead example

$$\Delta cp_t = \beta_b \Delta cp_{t-1} + \beta_f \mathbb{E}(\Delta cp_{t+1}|\Omega_t) + \gamma \Delta s_t + u_t, \tag{1}$$

where $\mathbb{E}(\Delta cp_{t+1}|\Omega_t)$ is the expectation made at time t of the future endogenous variable conditional on Ω_t, the information set available at time t. Indeed, we can substitute rational expectations $\mathbb{E}(\Delta cp_{t+1}|\Omega_t)$ either by a perfect foresight scheme $\mathbb{E}(\Delta cp_{t+1}|\Omega_t) = \Delta cp_{t+1}$ or by the

[1]Figure 2 of the empirical section displays the series we have used for our empirical investigation. The five monthly commodity price indices are the IMF primary commodity price indices and the US exchange rate is released by the Federal Reserve Bankof St. Louis. In Figure 2 we present the annual growth rates of the series, namely $(1 - L^{12})cp_{it}$, $i = 1, ..., 5$ and $(1 - L^{12})s_t$. Doing so we impose unit roots both at the zero and at each seasonal frequency.

sum of their realizations plus the realization of a martingale difference process $\mathbb{E}(\Delta cp_{t+1}|\Omega_t) = \Delta cp_{t+1} + \xi_{t+1}$ (see Broze, Gouriéroux and Szafarz, 1995). In both cases, this finally reduces to the MARX equation $(1 - \phi L)(1 - \varphi L^{-1})\Delta cp_t = \vartheta \Delta s_t + \varepsilon_t$ that we investigate in this article.

The remainder of the paper is organized as follows. Section 2 formalizes the notion of MARX models, their identifiability and how to simulate such processes. Section 3 considers (approximate) ML estimation and introduces a convenient way to compute standard errors which is not based on computing the Hessian using gradient-based (numerical) procedures. The results from various Monte Carlo simulations are collected in Section 4. Section 5 details the empirical applications. A careful reader of our model (1) will notice that our starting point is opposite to the common present value representation used to explain the link between commodity prices and exchange rates (Chen, Rogoff and Rossi, 2010). Indeed, using the quadratic determinantal equation method (Pesaran, 2015, 473-475), we can write (1) as

$$\Delta cp_t - \alpha_b \Delta cp_{t-1} = \left(\frac{\gamma}{1 - \beta_f \alpha_b}\right) \sum_{j=0}^{\infty} \alpha_f^{-1} \mathbb{E}(\Delta s_{t+j}|\Omega_t) + \left(\frac{1}{1 - \beta_f \alpha_b}\right) u_t, \qquad (2)$$

that is, a lag-augmented present value model where commodity prices Granger cause exchange rates as in Bork, Kaltwasser and Sercu (2014). Note that we do not consider commodity prices as any other financial asset depending on expectations only as past values of Δcp_t, which (probably) also embody values of fundamentals, that in turn also influence current price movements. Section 6 summarizes and concludes. Proofs and additional material can be found in the Appendix.

2 The MARX Model

Let y_t be the variable of interest which is observed over the time period $t = 1, ..., T$. Let x_i $(i = 1, ..., q)$ be a strictly exogenous variable for y_t and $\beta \in \mathbb{R}^q$ a vector of parameters. Then we can define $X_t = [x_{1,t}, ..., x_{q,t}]' \in \mathbb{R}^q$ as the vector of all exogenous variables[2] at time t. The MARX(r, s, q) for a stationary time series y_t can now be represented as

$$\phi(L)\varphi(L^{-1})y_t - \beta' X_t = \varepsilon_t, \tag{3}$$

where $\phi(L)$ is a lag polynomial of order r, $\varphi(L^{-1})$ a lead polynomial of order s and $r+s = p$. 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: $L^{-i} y_t = y_{t+i}$. The error term ε_t is assumed to be strong white noise. When $\varphi_1 = ... = \varphi_s = 0$, the process y_t is a purely causal autoregressive model with strictly exogenous variables, denoted MARX$(r, 0, q)$ or simply ARX(r, q):

$$\phi(L)y_t - \beta' X_t = \varepsilon_t. \tag{4}$$

Specification (4) can be seen as the standard backward-looking ARX model. Conversely, the process in (3) reduces to a purely noncausal MARX$(0, s, q)$:

$$\varphi(L^{-1})y_t - \beta' X_t = \varepsilon_t, \tag{5}$$

when $\phi_1 = ... = \phi_r = 0$. Note that the concepts of causality and noncausality are defined in terms of the strictly stationary solution of the model. To that end, we assume that both polynomials in (3) have their zeros outside the unit circle:

$$\phi(z) \neq 0 \text{ for } |z| \leq 1 \text{ and } \varphi(z) \neq 0 \text{ for } |z| \leq 1. \tag{6}$$

When $q = 0$, the process in (4) [(5)] reduces to a purely causal [noncausal] AR process that has a one-sided MA representation consisting of only past [future] and current values of ε_t. For

[2] We only consider contemporaneous values of X_t in the model. The MARX model can also take the form of a mixed autoregressive distributed lag (MARDL) model. See Appendix A for derivation and motivation.

the process in (3) these conditions however imply that the process y_t follows a two-sided MA representation involving past, current and future values of ε_t. In case $q > 0$, the processes considered no longer have a strictly stationary solution solely in terms of ε_t, but involve both X_t and ε_t. That is,

$$y_t = \pi(L, L^{-1})\varepsilon_t + \pi(L, L^{-1})\beta' X_t = \sum_{j=-\infty}^{\infty} \pi_j z_{t-j}, \tag{7}$$

where $z_{t-j} = \varepsilon_{t-j} + \sum_{i=1}^{q} \beta_i x_{i,t-j}$ and $\pi(L, L^{-1})$ is an operator satisfying $\pi(L, L^{-1})\phi(L)\varphi(L^{-1}) = 1$. Similar to Gouriéroux and Jasiak (2015), we note that the polynomials $\phi(L)$ and $\varphi(L^{-1})$ are invertible and their inverses create infinite series in L and L^{-1} respectively, causing (7) to hold almost surely (see Brockwell and Davis, 1991, proposition 13.3.1 for more details). We observe that y_t still has a two-sided MA-representation, but augmented with a second part involving linear combinations of past, current and future values of X_t. As for all i, $\beta_i x_{i,t}$ can be interpreted as a new series $\tilde{x}_{i,t}$ which is the old series $x_{i,t}$ multiplied by a constant term β_i, y_t in (7) can be seen to consist of two additive parts: (i) a two-sided MA 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 $[\phi(L)\varphi(L^{-1})]$ to $\pi(L, L^{-1})$.[3]

Lemma 1. *From* (3)*, we can construct the unobserved noncausal and causal components* (u, v) *similar to Lanne and Saikkonen (2011) and Gouriéroux and Jasiak (2015) and obtain:*

$$u_t \equiv \phi(L)y_t \leftrightarrow \varphi(L^{-1})u_t - \beta' X_t = \varepsilon_t, \tag{8}$$

$$v_t \equiv \varphi(L^{-1})y_t \leftrightarrow \phi(L)v_t - \beta' X_t = \varepsilon_t. \tag{9}$$

In order to ensure identifiability of the parameter vector β and to prove consistency of the

[3]The effects of two-sided linear filters, with a focus on seasonal adjustment, on the identification of mixed causal-noncausal models is studied in Hecq, Telg and Lieb (2016b).

ML estimator, we make the following assumptions:

Assumptions. The processes in X_t are assumed to be

(A1) zero mean, ergodic and (strictly) stationary;

(A2) strictly exogenous w.r.t. ε_t (X_t and ε_t are independent stochastic processes);

(A3) at most ℓ-dependent, for some $\ell < \infty$;

(A4) linearly independent.

2.1 Simulation of MARX Processes

The filtered values defined in Lemma 1 establish a deterministic dynamic relationship between the unobserved components (u_t, v_t), the exogenous variables X_t and the process y_t, which can be used to simulate various MARX(r, s, q) series. Gouriéroux and Jasiak (2015) show extensively how to simulate MAR(r, s) processes and make use of the independence of specific blocks of u, v and y. In this section, we extend their analysis to the MARX(r, s, q) case and show that the equivalence of different information sets still holds.

The main difficulty for generating MARX(r, s, q) with both $r, s \geq 1$ is the product of polynomials $\phi(L)\varphi(L^{-1})$. 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 (i.e., the purely causal, noncausal and static case), the problem is greatly simplified. We illustrate this by considering the MARX(0,1,1) model. In that case (3) reduces to:

$$y_t = \varphi_1 y_{t+1} + \beta_1 x_{1,t} + \varepsilon_t, \tag{10}$$

which can easily be simulated directly by generating a sequence of ε_t and choosing terminal values y_T and $x_{1,T}$.[4] In the general MARX(r, s, q) setup, filtered values are used to circumvent the problem. Defining $[\varphi(L^{-1})]^{-1} \equiv \delta(L^{-1})$, we can rewrite the second equality in (8) in the

[4]A burn-in period should be considered to delete dependence on terminal values.

following way:

$$u_t = \sum_{j=0}^{\infty} \delta_j \left(\sum_{i=1}^{q} \beta_i x_{i,t+j} + \varepsilon_{t+j} \right) = \sum_{j=0}^{\infty} \delta_j z_{t+j}. \tag{11}$$

In a similar fashion, when we take $[\phi(L)]^{-1} \equiv \delta(L)$, we obtain for v_t:

$$v_t = \sum_{j=0}^{\infty} \alpha_j \left(\sum_{i=1}^{q} \beta_i x_{i,t-j} + \varepsilon_{t-j} \right) = \sum_{j=0}^{\infty} \alpha_j z_{t-j}. \tag{12}$$

Using these expressions, MARX(r, s, q) can be constructed directly by means of the definitions given in (8) and (9). That is, the causal and noncausal components (u_t, v_t) can be simulated independently and can be interpreted as a causal [noncausal] "error term" of a purely noncausal [causal] autoregression.

Example 1. An MARX(1,1,1) can be simulated according to

$$y_t = \varphi_1 y_{t+1} + \sum_{j=0}^{\infty} \phi^j (\beta_1 x_{1,t-j} + \varepsilon_{t-j}),$$

where a truncation sufficiently large has to be considered for the infinite sum. Since $\deg(\phi(L)) = 1$ in this instance, it is straightforward to compute the inverse of this polynomial. For more complicated polynomials, one could compute a companion matrix to find its inverse.

Figure 1 shows simulated paths of an MAR(1,1) and MARX(1,1,1) process with $(\phi_1, \varphi_1) = (0.3, 0.5)$, $\beta = 0.3$, $x_{1,t} \overset{iid}{\sim} t(1)$ and $\varepsilon_t \overset{iid}{\sim} t(3)$. It can be seen that both processes generally move similarly with the major exception that the MARX process contains more peaks and troughs. This is due to the choice of $x_{1,t}$, which is chosen to be standard Cauchy distributed for expository purposes. Hence, the MARX specification takes into account shocks that cannot be explained by past, current and future values of the dependent variable, but which are present because of major changes in explanatory exogenous variables at some specific points in time. In order to justify the simulation method as outlined above, we present the following proposition that

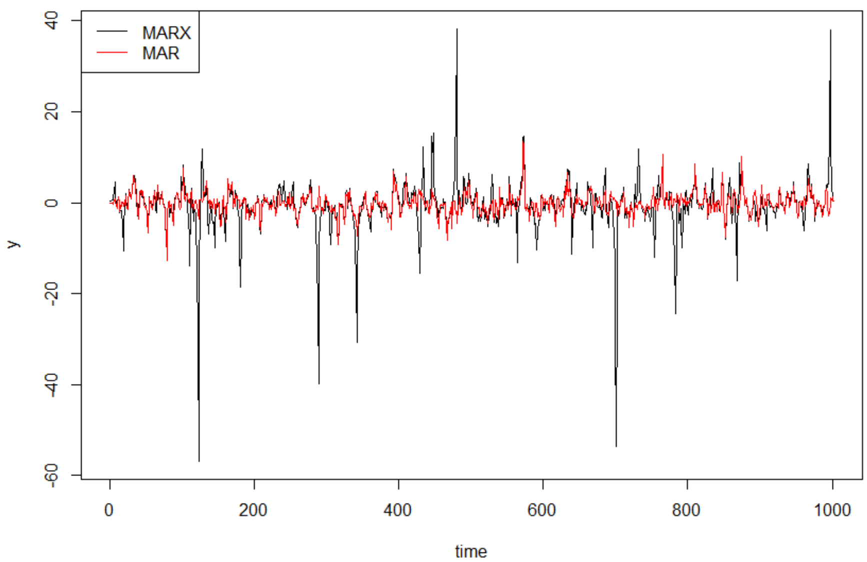

Figure 1: Simulated MAR(1,1) and MARX(1,1,1) process

shows the equivalence of various information sets. A proof of this proposition can be found in the Appendix (part B).

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

(*i*) $(y_1, ..., y_T, X_{r+1}, ..., X_{T-s})$

(*ii*) $(y_1, ..., y_r, u_{r+1}, ..., u_T, X_{r+1}, ..., X_{T-s})$

(*iii*) $(v_1, ..., v_{T-s}, y_{T-s+1}, ..., y_T, X_{r+1}, ..., X_{T-s})$

(*iv*) $(y_1, ..., y_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, u_{T-s+1}, ..., u_T)$

(*v*) $(v_1, ..., v_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, y_{T-s+1}, ..., y_T)$

(*vi*) $(v_1, ..., v_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, u_{T-s+1}, ..., u_T)$

Additionally, the following information sets are equivalent:

(i') $(y_1, ..., y_T)$

(ii') $(y_1, ..., y_r, u_{r+1}, ..., u_T)$

(iii') $(v_1, ..., v_{T-s}, y_{T-s+1}, ..., y_T)$

From this we deduce that the three sets of variables $(v_1, .., v_r)$, $(\varepsilon_{r+1}, ..., \varepsilon_{T-s})$ and $(u_{T-s+1}, ..., u_T)$ are also independent in the MARX setup. Similar to Gouriéroux and Jasiak (2015), we can still interpret $(v_1, .., v_r)$ $[(u_{T-s+1}, ..., u_T)]$ as the initial [terminal] conditions that determine the path of process y_t over the period $1, ..., T$.

2.2 Identifiability

Identifiability of mixed causal-noncausal models has received a lot of attention in the literature. Since the Gaussian distribution is fully characterized by its autocovariance function (which is symmetric), it is well-known that forward- and backward-looking behavior cannot be distinguished (see e.g., Breidt et al., 1991). For this reason, estimation methods based on solely second order properties of the data (like e.g., OLS and Gaussian MLE) cannot be used to identify such models. As the inclusion of exogenous variables introduces the presence of cross-covariances (which are generally not symmetric), MARX models can be identified even under Gaussianity.

To illustrate this, we consider the purely causal MARX(1,0,1) and the purely noncausal MARX(0,1,1) model and consider the kth-order autocovariance of y_t. Let $\gamma_y(k)$ denote the covariance between y_t and y_{t-k} and $\gamma_{xy}(s, t)$ the covariance between x_s and y_t. Then the autocovariance for the causal model can be written as

$$\gamma_y(k) = \phi_1 \gamma_y(k-1) + \beta_1 \gamma_{xy}(t, t-k),$$

and for the noncausal model as

$$\gamma_y(k) = \varphi_1 \gamma_y(k-1) + \beta_1 \gamma_{xy}(t-k, t).$$

When we have no exogenous variables, the second part on the right-hand side of both equations vanishes. The autocovariance at order k equals the respective autoregressive parameter times its autocovariance at $k-1$. In estimation methods like OLS, the minimization of the sum of squared residuals is based on solely this criterion and thus sets $\hat{\phi}_1 = \hat{\varphi}_1$, which means identification cannot be achieved. The cross-covariances $\gamma_{xy}(t, t-k)$ and $\gamma_{xy}(t-k, t)$ however need not be equal (and rarely are equal)[5], which causes them to create a different autocovariance structure for y_t depending on the usage of a causal, noncausal or mixed data generating process.

3 Maximum Likelihood Estimation

Maximum likelihood (ML) estimation of noncausal autoregressive models has been studied by Breidt, Davis, Lii and Rosenblatt (1991), Andrews, Breidt and Davis (2006) and Lanne and Saikkonen (2011).[6] They show that ML 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 strictly exogenous regressors.

Similar to Lanne and Saikkonen (2011), we shall assume that the density function $f(x; \lambda)$ satisfies the regularity conditions of Andrews et al. (2006).[7] The permissible parameter space of λ, denoted by Λ, is some subset of $\mathbb{R}^d$, $\sigma > 0$ and the permissible space of the parameters ϕ and

[5]Even when we assume joint stationarity between $x_{1,t}$ and y_t, the cross-covariances $\gamma_{xy}(k)$ and $\gamma_{xy}(-k)$ are generally unequal.

[6]Breidt 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 (3), the only difference being the exclusion of strictly exogenous variables X_t.

[7]The 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.

φ is defined by the stationarity condition (6). Using the independence of the blocks $(v_1, ..., v_r)$, $(\varepsilon_{r+1}, ..., \varepsilon_{T-s})$ and $(u_{T-s+1}, ..., u_T)$, it is shown in the Appendix (part C) that the density of the process y_t can be written as the product of the densities of these three variables. 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 (3) and taking logs, we obtain the following log-likelihood function:

$$\begin{aligned} L_T(\theta) &= \sum_{t=r+1}^{T-s} \ln f_\sigma(\phi(L)\varphi(L^{-1})y_t - \beta' X_t; \lambda) \\ &= \sum_{t=r+1}^{T-s} g_t(\theta), \end{aligned} \tag{13}$$

where $\theta = [\phi \ \varphi \ \beta \ \sigma \ \lambda]'$. 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 as in (8) and (9) to write the series u_t and v_t as functions of the parameters, i.e., $u_t(\phi)$ and $v_t(\varphi)$. Then we can characterize $g_t(\theta)$ as follows:

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

where we also used $f_\sigma(x; \lambda) = \sigma^{-1} f(\sigma^{-1} x; \lambda)$ (definition of density). Maximizing $L_T(\theta)$ over permissible values of θ gives an approximate ML estimator of θ. We assume for now that the orders r and s are known. Denote the true value of θ by θ_0 (and similarly for its components). Furthermore, assume that λ_0, the true value of λ, is an interior point of Λ.

3.1 Asymptotic Properties of the Approximate ML Estimator

We first consider the score of θ evaluated at true parameter values. Define $V_{t-1} = (v_{t-1}, ..., v_{t-r})$ and $U_{t+1} = (u_{t+1}, ..., u_{t+s})$, where u_t and v_t are defined in terms of true parameter values, i.e., $u_t = \sum_{j=0}^{\infty} \delta_{0j} \left(\sum_{i=1}^{q} \beta_{0i} x_{i,t+j} + \varepsilon_{t+j} \right)$ and $v_t = \sum_{j=0}^{\infty} \alpha_{0j} \left(\sum_{i=1}^{q} \beta_{0i} x_{i,t-j} + \varepsilon_{t-j} \right)$. By direct differentation of (13), we obtain:

$$\frac{\partial}{\partial \phi} g_t(\theta_0) = -\frac{f'(\sigma_0^{-1}\varepsilon_t; \lambda_0)}{\sigma_0 f(\sigma_0^{-1}\varepsilon_t; \lambda_0)} V_{t-1},$$

$$\frac{\partial}{\partial \varphi} g_t(\theta_0) = -\frac{f'(\sigma_0^{-1}\varepsilon_t; \lambda_0)}{\sigma_0 f(\sigma_0^{-1}\varepsilon_t; \lambda_0)} U_{t+1},$$

and

$$\frac{\partial}{\partial \beta} g_t(\theta_0) = -\frac{f'(\sigma_0^{-1}\varepsilon_t; \lambda_0)}{\sigma_0 f(\sigma_0^{-1}\varepsilon_t; \lambda_0)} X_t,$$

where $f'(x; \lambda) = \partial f(x; \lambda)/\partial x$ and use has been made of the fact that $\varphi_0(L^{-1})u_t - \beta_0' X_t = \varepsilon_t = \phi_0(L)v_t - \beta_0' X_t$. Similarly, for the distributional parameters:

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

and

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

The following lemma presents the asymptotic distribution of the score vector. Define

$$\mathcal{J} = \int \frac{(f'(x; \lambda_0))^2}{f(x; \lambda_0)} dx > 1 \quad \text{and} \quad \mathcal{I} = \int x^2 \frac{(f'(x; \lambda_0))^2}{f(x; \lambda_0)} dx - 1,$$

where the first inequality follows from Remark 2 in Andrews et al. (2006). Furthermore set

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

The matrix Σ is symmetric and has the matrices $\Sigma_{11} = \sigma_0^{-2}\mathcal{J}\Gamma_V, \Sigma_{22} = \sigma_0^{-2}\mathcal{J}\Gamma_U, \Sigma_{33} = \sigma_0^{-2}\mathcal{J}\Gamma_X$ on the diagonal, where Γ_V and Γ_U are the autocovariance matrices of V_{t-1} and U_{t+1} respectively. Γ_X is the cross-covariance matrix of the q processes in X_t. Σ_{12} is a $(r \times s)$ matrix where the (i, j)th element equals:

$$\sum_{t=0}^{\infty} \alpha_t \delta_{t+i-j} + \sigma^{-2}\mathcal{J} \sum_{a=0}^{\infty}\sum_{b=0}^{\infty} \alpha_a \delta_b \sum_{m=1}^{q}\sum_{l=1}^{q} \beta_m \beta_l \gamma_{x_m, x_l}(i + j + a + b).$$

The Σ_{13} matrix has size $(r \times q)$ with the (i, j)th element equal to

$$\sigma^{-2}\mathcal{J} \sum_{a=0}^{\infty} \alpha_a \sum_{l=1}^{q} \beta_l \gamma_{x_l, x_j}(i + a),$$

while for Σ_{23} this element is

$$\sigma^{-2}\mathcal{J} \sum_{b=0}^{\infty} \delta_b \sum_{l=1}^{q} \beta_l \gamma_{x_l, x_j}(i + b).$$

Note that the cross-covariance involves 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 $(= 1, ..., q)$ runs over all exogenous variables $x_{1,t}, ..., x_{q,t}$. Finally define the $(d + 1) \times (d + 1)$ matrix

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

where

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

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

and

$$\omega_\sigma^2 = \omega_0^{-2}\mathcal{I}.$$

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

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

Moreover, the matrices Σ and Ω 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, Davis and Breidt (2006). This ensures that both the autoregressive and exogenous (variables) parameters are orthogonal to the distributional parameters. The positive definiteness of Ω is assumed through condition (A6) of Andrews et al. (2006). The positive definiteness of Σ follows from $\mathcal{J} > 1$.

Theorem 1. *If conditions (A1)-(A7) of Andrews et al. (2006) hold, there exists a sequence of local maximizers $\hat{\theta} = [\hat{\phi}\ \hat{\varphi}\ \hat{\beta}\ \hat{\sigma}\ \hat{\lambda}]'$ of $L_T(\theta)$ in* (13) *such that*

$$\sqrt{n}(\theta_{ML} - \theta_0) \xrightarrow{d} N(0, diag(\Sigma^{-1}, \Omega^{-1})).$$

3.2 Computing the Covariance Matrix

Block diagonality of the covariance matrix in Theorem 1 shows that the approximate ML estimators of the model parameters $[\phi\ \varphi\ \beta]'$ and the distributional parameters $[\sigma\ \lambda]'$ are asymptotically independent. The computation of this covariance matrix 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 autoregressive parameters for Student's t-MLE and the least absolute deviations (LAD) estimator.[8]

3.2.1 Student's t Maximum Likelihood Estimation

Similar to Hecq, Lieb and Telg (2016), we can characterize the asymptotic distribution of the Student's t-MLE and LAD estimated parameters in the finite variance framework. If $\nu > 2$, the MLE is $\sqrt{n}$-consistent and asymptotically normal. Define the $(1 \times n)$ series $u \equiv U_t^* = (u_1, ..., u_{T-s})$ up to $U_{t+s}^* = (u_{s+1}, ..., u_T)$, $V_{t-r}^* = (v_1, ..., v_{T-r})$ up to $v \equiv V_t^* = (v_{r+1}, ..., v_T)$, $X_{i,t}^u = (x_{i,1}, ..., x_{i,T-s})$ and $X_{i,t}^v = (x_{i,r+1}, ..., x_{i,T})$. Then we construct the matrices $Z = [U_{t+1}^*, ..., U_{t+s}^*, X_{1,t}^u, ..., X_{q,t}^u]'$ and $Q = [V_{t-1}^*, ..., V_{t-r}^*, X_{1,t}^v, ..., X_{q,t}^v]'$, which are of dimensions $((s+q) \times n)$ and $((r+q) \times n)$ respectively. Using this notation, we can write the autoregressions defined in (8) and (9) in matrix notation as follows:

$$u = \zeta' Z + \varepsilon, \tag{14}$$

$$v = \xi' Q + \varepsilon, \tag{15}$$

with $\zeta = [\varphi \ \beta]' \in \mathbb{R}^{s+q}$ and $\xi = [\phi \ \beta]' \in \mathbb{R}^{r+q}$.

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

$$\sqrt{n}(\zeta_{ML} - \zeta_0) \xrightarrow{d} N\left(0, \frac{\nu+3}{\nu+1}\sigma^2 \Upsilon_\phi^{-1}\right), \tag{16}$$

[8] The LAD estimator can be interpreted as a maximum likelihood estimator for which the error term follows a Laplacian distribution. It should be noted that the density of this distribution does not satisfy the regularity conditions of Andrews et al. (2006). Since the LAD can be used as an initial estimator for ϕ and φ (Lanne and Saikkonen, 2011) and is found to outperform Student's t-MLE in certain instances (Hecq et al., 2016), we also establish the asymptotic distribution of the model parameters for the LAD estimator.

$$\sqrt{n}(\xi_{ML} - \xi_0) \xrightarrow{d} N\left(0, \frac{\nu+3}{\nu+1}\sigma^2 \Upsilon_{\varphi}^{-1}\right), \tag{17}$$

holds. We use the notation $\Upsilon_{\boldsymbol{\varphi}} = \mathbb{E}[QQ']$ and $\Upsilon_{\boldsymbol{\phi}} = \mathbb{E}[ZZ']$, where $\boldsymbol{\varphi}$ and $\boldsymbol{\phi}$ signify the relation between the unobserved values u,v and y as defined in (8)-(9). These quantities can be estimated consistently by $(1/n)\sum_{i=1}^{n} Q_i Q_i'$ and $(1/n)\sum_{i=1}^{n} Z_i Z_i'$, where Q_i [resp. Z_i] denotes the ith column of the matrix Q [resp. Z]. For large ν, i.e., $\nu \to \infty$, l_y approaches the Gaussian (log)-likelihood, and the model parameters cannot be consistently estimated anymore.

3.2.2 Least Absolute Deviation Estimation

Using the model specifications in (14)-(15), we can derive a similar result for the LAD estimator. Let λ denote a vector of distributional parameters. If ε_t is a sequence of *iid* random variables with mean zero, median zero, finite variance and probability density function $f_\varepsilon(\varepsilon_t; \lambda)$ that is continuous in a neighborhood of zero, the LAD estimator is $\sqrt{n}$-consistent and asymptotically normal (Wu and Davis, 2010). Following Hecq et al. (2016), it follows that, conditional on the unobserved causal and noncausal components,

$$\sqrt{T}(\zeta_{LAD} - \zeta_0) \xrightarrow{d} N\left(0, \frac{1}{4f_\varepsilon^2(0)} \Upsilon_{\phi}^{-1}\right), \tag{18}$$

$$\sqrt{T}(\xi_{LAD} - \xi_0) \xrightarrow{d} N\left(0, \frac{1}{4f_\varepsilon^2(0)} \Upsilon_{\varphi}^{-1}\right), \tag{19}$$

where $f_\varepsilon(0)$ can be estimated by a logistic kernel.

4 Simulation Study

By means of Monte Carlo simulations, we investigate three different cases of interest: (i) the performance of the MLE for MARX processes, (ii) the identifiability of MARX models under Gaussianity and (iii) a model selection procedure for MARX models. Each table reports results for 1000 replications.

4.1 Performance MLE for MARX

To assess the performance of the maximum likelihood estimator, we take the following MARX(1,1,1) as data generating process (DGP):

$$(1 - \phi_1 L)(1 - \varphi_1 L^{-1})y_t - \beta_1 x_{1,t} = \varepsilon_t, \tag{20}$$

where $\phi_1 = 0.3$, $\varphi_1 = 0.5$ and $\beta_1 = 0.3$. The error term ε_t follows a t-distribution with 2 degrees of freedom. $x_{1,t}$ will follow different specifications:

(1) $x_{1,t} \overset{iid}{\sim} t(5)$,

(2) $x_{1,t} \overset{iid}{\sim} N(0,1)$,

(3) $x_{1,t} \overset{iid}{\sim} C(0,1)$,

(4) $x_{1,t}$ follows an AR(1) process: $x_{1,t} = 0.6x_{1,t-1} + \epsilon_t$ where $\epsilon_t \overset{iid}{\sim} N(0,5)$.

Table 1 reports the mean and standard deviations of the estimated parameters by MLE over all simulations. It can be seen that different specifications for $x_{1,t}$ only introduce relatively small differences. Most noticeably, 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 $t(5)$ and $N(0,1)$ distribution do not generate vast outliers in $x_{1,t}$, which makes 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 very close. For all four specifications, the most difficult parameter to estimate is ν, which has a very large standard deviation for $T = 50$. For larger T, the standard deviations decrease rapidly. In all cases, the estimated mean over all parameters becomes more accurate and standard deviations decline as T grows large, which suggests the consistency of the maximum likelihood estimator in this framework.

		Specification for $x_{1,t}$							
		$x_{1,t} \overset{iid}{\sim} t(5)$		$x_{1,t} \overset{iid}{\sim} N(0,1)$		$x_{1,t} \overset{iid}{\sim} C(0,1)$		$x_{1,t} \sim$ AR(1)	
T	Parameter	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
50	ϕ_1	0.299	0.120	0.303	0.139	0.299	0.081	0.292	0.071
	φ_1	0.481	0.117	0.480	0.138	0.494	0.072	0.494	0.060
	β_1	0.308	0.167	0.313	0.205	0.299	0.043	0.303	0.037
	ν	2.615	4.446	2.594	2.932	2.490	2.470	2.459	3.196
100	ϕ_1	0.298	0.075	0.301	0.071	0.299	0.035	0.295	0.048
	φ_1	0.498	0.067	0.492	0.065	0.500	0.031	0.495	0.041
	β_1	0.300	0.110	0.301	0.139	0.299	0.021	0.302	0.026
	ν	2.129	0.565	2.150	0.533	2.132	0.535	2.173	0.600
500	ϕ_1	0.300	0.024	0.299	0.025	0.300	0.012	0.299	0.018
	φ_1	0.500	0.021	0.500	0.021	0.500	0.007	0.500	0.016
	β_1	0.301	0.046	0.298	0.058	0.300	0.004	0.300	0.010
	ν	2.034	0.187	2.027	0.185	2.012	0.221	2.026	0.190
1000	ϕ_1	0.300	0.015	0.299	0.015	0.300	0.005	0.300	0.012
	φ_1	0.500	0.014	0.500	0.013	0.500	0.005	0.500	0.010
	β	0.300	0.032	0.299	0.041	0.300	0.002	0.300	0.007
	ν	2.007	0.128	2.016	0.136	2.000	0.153	2.014	0.131

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

4.2 Are MARX Models Identifiable Under Gaussianity?

In Section 2.2, we discussed the identifiability of MARX models even when the error term is normally distributed. To evaluate this important feature, we consider the purely noncausal data generating process as defined in (10) with both ε_t and $x_{1,t} \overset{iid}{\sim} N(0,1)$, with various combinations of parameter values for φ_1 and β_1. To the simulated series y_t we fit both a correctly specified model, i.e., the MARX(0,1,1) and a misspecified autoregressive model with exogenous regressors, i.e. the MARX(1,0,1). The estimation method used is Gaussian MLE.[9] We select the model

[9]Due to condition (A5) of Andrews et al. (2006), consistency of Gaussian MLE for the MARX was not shown in Section 3 of this paper. However, the consistency of Gaussian MLE for the pure ARX case is established in Hannan, Dunsmuir and Deistler (1980).

that has the highest value for the log-likelihood function at the estimated parameters. Table 2 shows the percentages with which the correct model is chosen for different parameter values for φ_1 and β_1.

We observe that the correct model is selected in approximately 50% of the cases when $\beta_1 = 0$ (irrespective of the value for φ_1). This is exactly in line with our expectations, as purely causal, mixed and purely noncausal autoregressive models cannot be identified by Gaussian MLE. Another special case arises when $\varphi_1 = 0$, as their is no autoregressive part present in the model. When $\beta_1 = 0$, we have a strong white noise and thus the correct model specification is not among the options. This is also the case for $\beta_1 \neq 0$, which causes the DGP to become a static regression. In both instances, both the MARX(1,0,1) and MARX(0,1,1) are chosen with roughly equal frequencies.

φ_1	β_1	$T = 50$	$T = 100$	$T = 200$	$T = 500$
0	0	49.0	51.9	49.1	50.0
	0.3	49.6	51.1	49.8	49.4
	0.7	49.4	50.5	49.3	49.1
0.1	0	50.1	51.9	49.2	50.0
	0.3	50.9	54.6	57.8	63.0
	0.7	55.5	61.3	68.6	82.7
0.4	0	52.2	52.4	50.3	50.8
	0.3	67.3	77.5	85.7	96.7
	0.7	88.7	96.6	99.6	100.0
0.6	0	53.4	53.0	50.4	51.3
	0.3	74.3	86.9	94.9	99.8
	0.7	96.0	99.7	100.0	100.0
0.8	0	50.9	50.2	48.6	51.4
	0.3	79.3	92.0	97.9	100.0
	0.7	98.9	100.0	100.0	100.0
0.9	0	50.4	48.9	47.5	51.9
	0.3	78.4	92.7	98.8	100.0
	0.7	99.1	99.9	100.0	100.0

Table 2: Frequency (in %) with which the correct MARX(0,1,1) model is selected

For $\varphi_1 \neq 0$, we see the same pattern in every case: identification of the correct model increases with β_1 and with sample size T. For a higher value of β, the cross-covariance term becomes more important in determining the autocovariance of y_t, which is different for a purely causal and purely noncausal MARX. Because of this, Gaussian MLE is able to distinguish between the two specifications in contrast to the case without exogenous regressors. In the same spirit, Cubadda, Hecq, Lieb and Telg (2016) show that reduced rank restrictions help to identify purely causal from purely noncausal VAR models in a Gaussian framework whareas unrestricted models are not identifiable by Gaussian MLE.

4.3 Model Selection

Lanne and Saikkonen (2011) propose a two-step approach to perform model selection for mixed causal-noncausal models MAR(r, s). In a first step, purely causal autoregressive processes are estimated by OLS or Gaussian MLE and the lag order p is determined by conventional information criteria like AIC, BIC and HQ.[10] As soon as p is identified, one selects a model among all MAR(r, s) with $p = r + s$. The model that attains the highest value for the log-likelihood at its estimated parameters is chosen to be the final model. This simulation study checks whether both steps are still valid in the MARX framework. To that end, we simulate (20) with $\phi_1 = 0.3$, $\varphi_1 = 0.5$ and $\beta_1 = 0.3$, $\varepsilon_t \overset{iid}{\sim} t(3)$ and $x_{1,t} \overset{iid}{\sim} t(2)$. Purely causal ARX$(p, 1)$ models are estimated by Gaussian MLE, where $p = 0, ..., 4$. Table 3, top part, shows the percentages with which AIC, BIC and HQ select a certain order p (true order equals 2). As comparison, the numbers within parentheses are the corresponding frequencies for the MAR model based on the same specifications only without exogenous variable $x_{1,t}$, i.e. $\beta_1 = 0$.

We can see that all information criteria tend to underestimate the true lag order by one (especially BIC). 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

[10] In empirical work, it is advised to perform diagnostic tests to see whether additional lags are needed to remove autocorrelation from the series. Also a normality test on the residuals might be performed to test for signs of noncausality. A description of the model selection procedure can be found in Hecq et al. (2016).

	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
	(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	61.6	82.4	72.6	55.3	74.8	65.3	30.0	47.0	34.9	0.2	4.1	0.7
	(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	33.5	16.4	25.2	41.8	24.1	32.9	67.1	52.9	63.8	96.4	95.5	97.6
	(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	4.6	0.9	1.9	2.6	0.8	1.6	2.9	0.1	1.3	3.4	0.4	1.7
	(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
	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
	Lag-lead order (r, s) selected by highest log-likelihood with known p=2											
	(2,0)	(1,1)	(0,2)	(2,0)	(1,1)	(0,2)	(2,0)	(1,1)	(0,2)	(2,0)	(1,1)	(0,2)
	35.5	54.4	10.1	29.0	68.8	2.2	16.9	83.0	0.1	6.8	93.2	0.0
	(3.3)	(78.7)	(18.0)	(0.2)	(95.4)	(4.4)	(0.0)	(99.9)	(0.1)	(0.0)	(100.0)	(0.0)

Table 3: Frequency (in %) with which the autoregressive orders p and (r, s) are selected

well-known that information criteria are derived from asymptotic properties and thus 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 (number in parentheses). This stresses the usage of diagnostic tests to discover model fit improvements. Similar to Lanne and Saikkonen (2011), we find that the choice of parameter values influences the frequency with which a certain model is selected. For example, when $p = 2$, a very low causal parameter value and a very high noncausal parameter value leads to the selection of a first-order noncausal process (and vice versa).

In a second exercise, we suppose the correct order $p = 2$ is known and investigate the selection of MARX$(r, s, 1)$ models with $r + s = 2$ based on the highest log-likelihood. In Table 3, bottom part, we observe that the model selection procedure improves with sample size, finding the correct specification in more than 80% of the cases for $T = 500$ and more than 90% when $T = 1000$. As data sets in empirical studies are often smaller, we advise practitioners to interpret the results with caution. We find that the correct model is only selected in a little more than half of the cases when $T = 100$, which suggests the use of complementary analysis (e.g., bootstrap or cross-validation criteria). In comparison, the MAR model selection is a lot more precise, as the correct model is already chosen in roughly 80% of the cases when $T = 100$.

5 Empirical Application

5.1 The data

We consider non seasonally adjusted monthly commodity prices $CP_{i,t}$ from 1980:01 to 2016:10, i.e., 442 observations for $i = 1...5$ indexes released by the IMF.[11] These are benchmark prices which are representative of the global market. They are determined by the largest exporter of a given commodity. IMF releases many different individual commodity prices but we only focus on the following indexes for this study:

- BEVE: Beverage Price Index, 2005 = 100, includes Coffee, Tea, and Cocoa,
- INDU: Industrial Inputs Price Index, 2005 = 100, includes Agricultural Raw Materials and Metals Price Indices,
- RAWM: Agricultural Raw Materials Index, 2005 = 100, includes Timber, Cotton, Wool, Rubber, and Hides Price Indices,
- META Metals Price Index, 2005 = 100, includes Copper, Aluminum, Iron Ore, Tin, Nickel, Zinc, Lead, and Uranium Price Indices,
- OIL: Crude Oil (petroleum), Price index, 2005 = 100, simple average of three spot prices; Dated Brent, West Texas Intermediate, and the Dubai Fateh.

We also consider for the same period S_t, the trade weighted U.S. dollar index: broad, index Jan 1997=100, monthly, not seasonally adjusted. This series is taken from the Federal Reserve Bank of St. Louis database.[12] As commodities are mainly priced in dollar we can expect a contemporaneous relation between commodity prices and the US exchange rate.

[11] IMF Primary Commodity Prices, see http://www.imf.org/external/np/res/commod/index.aspx.
[12] Series name TWEXBMTH at https://fred.stlouisfed.org.

We do not reject a unit root at the zero frequency in each series. However first differences of the logs (monthly growth rates) are quite volatile and some seasonal components are still present. We do not formally test for monthly unit root tests but decide to impose them. We consequently work with annual growth rates $(1 - L^{12}) \ln CP_{i,t} = \Delta_{12} cp_{i,t}$ and $(1 - L^{12}) \ln S_t = \Delta_{12} s_t$. Series are displayed in Figure 2. Both commodity indices and exchange rates being stock data this step has the additional advantage to allow the comparison of series selected a different frequencies without changing the definition of the series: annual growth rates can be sampled monthly, quarterly or annually.

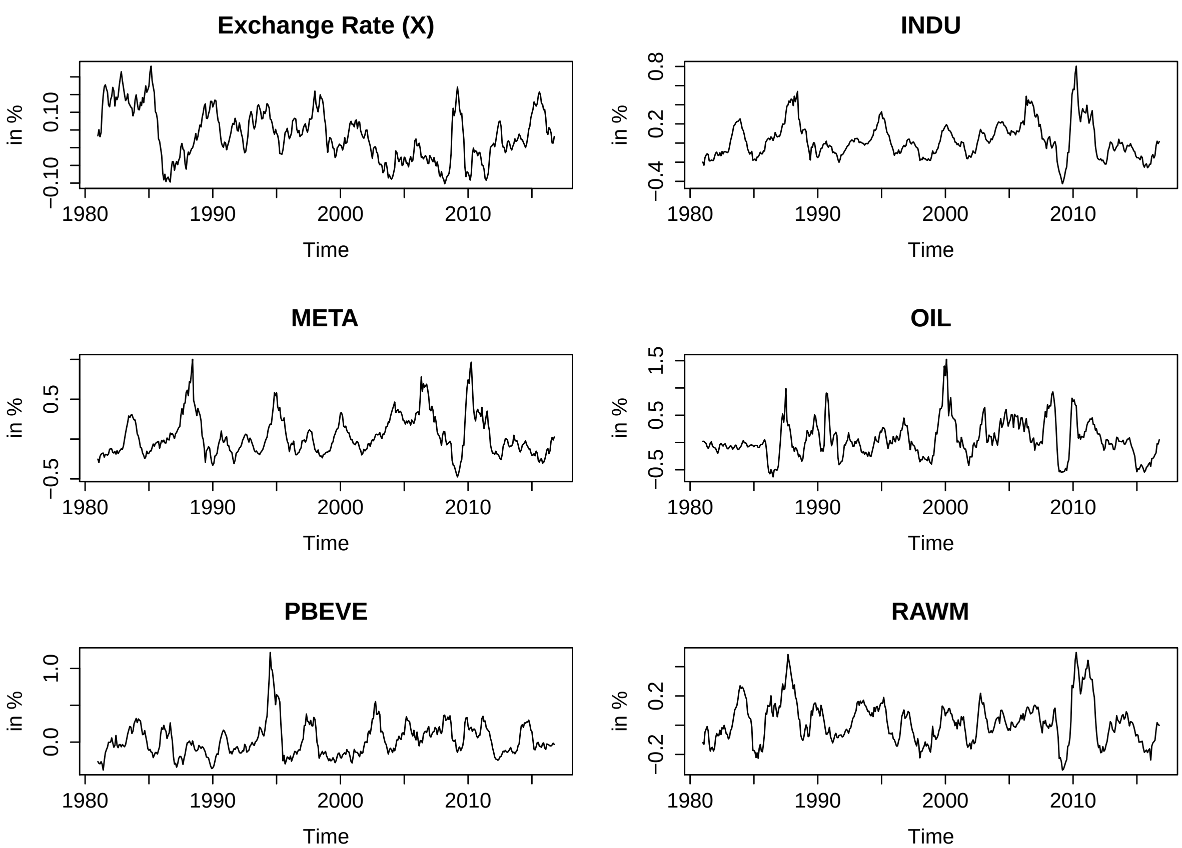

Figure 2: Growth rates of the commodity prices and exchange rate

5.2 From RE to MARX models

Lanne and Luoto (2013) directly link mixed causal-noncausal models to the analysis of inflation using the hybrid NKPC. We can do the same with our relationship (1) in the introductory section that we write in annual differences

$$\Delta_{12}cp_t = \beta_b \Delta_{12}cp_{t-1} + \beta_f \mathbb{E}(\Delta_{12}cp_{t+1}|\Omega_t) + \gamma\Delta_{12}s_t + u_t. \tag{21}$$

By means of replacing expectations by the future realized value of inflation plus a martingale difference ξ_{t+1} and putting the three parts $(\mathbb{E}(\Delta_{12}cp_{t+1}|\Omega_t) - \xi_{t+1})$, $\Delta_{12}s_t$ and u_t into the error term, they show that one obtains an MAR(1,1) with a newly defined disturbance term, say η_t, that is likely to be autocorrelated. The variable $\Delta_{12}s_t$ is however likely to create autocorrelation and by assuming an MAR$(r-1, s-1)$ structure on η_t, they show that the process for $\Delta_{12}cp_t$ can be rewritten as a causal polynomial $\phi(L)$ of order r and a noncausal polynomial $\varphi(L^{-1})$ of order s. The information resulting from $\Delta_{12}s_t$ is lost however in this mixed causal-noncausal for $\Delta_{12}cp_t$. We prefer adding and subtracting $\beta_f\Delta_{12}cp_{t+1}$ and rearranging terms, which gives

$$\Delta_{12}cp_t = \beta_b\Delta_{12}cp_{t-1} + \beta_f\Delta_{12}cp_{t+1} + \gamma\Delta_{12}s_t + \underbrace{u_t + \beta_f\left(\mathbb{E}(\Delta_{12}cp_{t+1}|\Omega_t) - \xi_{t+1}\right)}_{\omega_{t+1}}. \tag{22}$$

Now we lag (22) by one period and subsequently divide this equation by β_f to obtain

$$\beta_f^{-1}\Delta_{12}cp_{t-1} = \Delta_{12}cp_t + \beta_b\beta_f^{-1}\Delta_{12}cp_{t-2} + \gamma\beta_f^{-1}\Delta_{12}s_{t-1} + \beta_f^{-1}\omega_t,$$

which can be rewritten in the following way

$$(1 - \beta_f^{-1}L + \beta_b\beta_f^{-1}L^2)\Delta_{12}cp_t = -\gamma\beta_f^{-1}\Delta_{12}s_{t-1} + \beta_f^{-1}\omega_t. \tag{23}$$

We want to write $a(L) = (1 - \beta_f^{-1}L + \beta_b\beta_f^{-1}L^2)$ as $a(L) = (1 - \phi L)(1 - \varphi^* L)$, where $|\phi| < 1$ and $|\varphi^*| > 1$. That is, we split the original polynomial in two different ones: one having all roots outside the unit circle $[\phi(z)]$ and one having its roots inside the unit circle $[\varphi^*(z)]$. With plausible values of α and γ, this can be done by taking

$$\phi = \frac{1}{2}\left(\beta_f^{-1} - \sqrt{\beta_f^{-2} - 4\beta_f^{-1}\gamma}\right) \text{ and } \varphi = \frac{1}{2}\left(\beta_f^{-1} + \sqrt{\beta_f^{-2} - 4\beta_f^{-1}\gamma}\right),$$

as was shown in Lanne and Luoto (2013). Following Lanne and Saikkonen (2011), we can rewrite the polynomial with roots inside the unit circle as a polynomial in reverse time with roots outside the unit circle. That is,

$$\begin{aligned}(1 - \phi z)(1 - \varphi^* z) &= (1 - \phi z)[-\varphi^* z(-\frac{1}{\varphi^*}z^{-1} + 1)] \\ &= -\varphi^* z(1 - \phi z)(1 - \varphi z^{-1}),\end{aligned}$$

where $\varphi = \frac{1}{\varphi^*}$. This result can be substituted in for the polynomial $a(L)$ in (23) to obtain

$$-\varphi^* L(1 - \phi L)(1 - \varphi L^{-1})\Delta_{12}cp_t = -\gamma\beta_f^{-1}\Delta_{12}s_{t-1} + \beta_f^{-1}\omega_t,$$

which by rearranging terms, reduces to a mixed causal-noncausal model, i.e.,

$$(1 - \phi L)(1 - \varphi L^{-1})\Delta_{12}cp_t = \underbrace{\gamma(\varphi^*\beta_f)^{-1}}_{\vartheta}\Delta_{12}s_t + \underbrace{(\varphi^*\beta_f)^{-1}\omega_{t+1}}_{\varepsilon_t}. \tag{24}$$

Since ω_{t+1} is assumed to be *iid* in time, also ε_t is *iid* in time. Hence, we obtain the MARX(1,1,1) model. Note that we consider a simple one-lag one-lead example in (1) from the outset. Hence, analysis can be extended to accommodate higher order processes.

5.3 Identification of MARX

The first step of our modeling strategy consists in fitting for each $(1 - L^{12}) \ln CP_{i,t} = \Delta_{12} cp_{i,t}$ series an OLS regression (or by Gaussian MLE) on their own lags and the contemporaneous value of $\Delta_{12} s_t$. The first part of Table 4 reports that step and provides the results concerning models chosen by different information criteria. Given the simulation results of Section 4, we decide to rely on HQ.

The second part of Table 4 reports those estimation results (standard errors in brackets). We observe that commodity prices depend on their own lags as well as on the exchange rate (but BEVE). The highest negative effect is on the OIL index, a result that makes sense given that oil products heavily depend on exports and hence are negatively influenced by an increase of the USD. The last row of Table 4 reports the value of the Jarque-Bera normality test. It is observed that we reject the null of normality in every equation and hence that we will be potentially able to discriminate components of the MARX model using a non-Gaussian MLE. Note that the possibility to identify models by Gaussian MLE or OLS using a strictly exogenous regressor is a feature of pure models (either causal or noncausal) that does not extend to mixed processes. A non-Gaussian MLE is needed for the latter case.

Once the number of lags p_i in pure autoregressive models with the contemporaneous $\Delta_{12} s_t$ is determined for each commodity price ($i = 1...5$), we estimate every possible MARX$(r_i, s_i, 1)$ models with $p_i = r_i + s_i$. We choose the model that gives the highest log-likelihood values. Table 5 reports the final results for each commodity. The reported values for ϕ (resp. φ) are the estimated coefficients of the lag (resp. lead) polynomials. We can observe mixed models in the five cases with some differences in the dynamics. Distributions are rather leptokurtic: the smallest value is $\hat{\nu} = 2.68$ for OIL, the largest value is $\hat{\nu} = 5.59$ for RAWM. The negative impact of exchange rate is more pronounced for OIL and the smallest value (significant now at 10%) is for BEVE index.

	Commodities				
	INDU	META	OIL	BEVE	RAWM
p_{AIC}	6	6	4	8	7
p_{BIC}	3	6	4	2	4
p_{HQ}	6	6	4	8	4
c	0.009	0.013	0.018	0.003	0.006
	(0.003)	(0.004)	(0.007)	(0.004)	(0.002)
a_1	1.190	1.134	1.201	1.206	1.161
	(0.049)	(0.048)	(0.048)	(0.049)	(0.048)
a_2	-0.225	-0.203	-0.401	-0.217	-0.170
	(0.075)	(0.072)	(0.075)	(0.077)	(0.074)
a_3	0.048	0.114	0.253	0.046	0.044
	(0.076)	(0.072)	(0.075)	(0.076)	(0.074)
a_4	-0.134	-0.220	-0.194	0.066	-0.135
	(0.076)	(0.072)	(0.047)	(0.075)	(0.046)
a_5	0.177	0.287		-0.250	
	(0.075)	(0.720)		(0.075)	
a_6	-0.159	-0.221		0.243	
	(0.047)	(0.047)		(0.076)	
a_7				0.037	
				(0.076)	
a_8				-0.144	
				(0.048)	
β_1	-0.148	-0.210	-0.286	-0.020	-0.101
	(0.036)	(0.053)	(0.087)	(0.046)	(0.030)
$\bar{R}^2$	0.94	0.92	0.87	0.91	0.92
JB	341.76	576.43	235.48	208.59	39.23

Table 4: Estimation results - pseudo causal models

	MARX(3,3,1)					
INDU	ϕ_1	0.268 (0.036)	φ_1	0.908 (0.035)	β_1	-0.100 (0.020)
	ϕ_2	0.223 (0.036)	φ_2	-0.175 (0.049)	c	0.004 (0.002)
	ϕ_3	0.029 (0.035)	φ_3	0.142 (0.035)	$[\nu, \sigma]$	[3.082, 0.028]
	MARX(5,1,1)					
META	ϕ_1	0.312 (0.028)	ϕ_4	-0.070 (0.044)	β_1	-0.148 (0.029)
	ϕ_2	0.081 (0.044)	ϕ_5	0.179 (0.029)	c	0.002 (0.002)
	ϕ_3	0.127 (0.045)	φ_1	0.794 (0.018)	$[\nu, \sigma]$	[2.960, 0.039]
	MARX(3,1,1)					
OIL	ϕ_1	0.466 (0.031)	φ_1	0.689 (0.024)	β_1	-0.211 (0.052)
	ϕ_2	0.020 (0.047)			c	0.008 (0.004)
	ϕ_3	0.176 (0.031)			$[\nu, \sigma]$	[2.679, 0.070]
	MARX(7,1,1)					
BEVE	ϕ_1	0.656 (0.038)	ϕ_5	-0.188 (0.059)	β_1	-0.059 (0.033)
	ϕ_2	0.042 (0.038)	ϕ_6	0.139 (0.059)	c	0.001 (0.003)
	ϕ_3	0.167 (0.059)	ϕ_7	-0.060 (0.037)	$[\nu, \sigma]$	[4.053, 0.047]
	ϕ_4	0.060 (0.059)	φ_1	0.540 (0.032)		
	MARX(2,2,1)					
RAWM	ϕ_1	0.716 (0.043)	φ_2	-0.024 (0.043)	β_1	-0.060 (0.022)
	ϕ_2	0.142 (0.043)			c	0.003 (0.002)
	φ_1	0.451 (0.043)			$[\nu, \sigma]$	[5.595, 0.033]

Table 5: Estimation results - MARX models

6 Conclusion

This paper proposes to estimate mixed causal-noncausal models when additional regressors are present. We have in mind the estimation of structural relationships subject to rational expectation schemes such as the new Hybrid Phillips curve or lag-augmented present value models. We provide a successful empirical illustration on the relation between commodity prices and the USD.

Our one step approach to estimating MARX is easy to implement[13] and the estimation of the standard errors that we propose is quite robust to computation overflows. It allows to estimate directly the impact of an exogenous variable without the need to augment the MAR with leads and lags (and to lose the impact of X_t) or to use a two-step approach as in Lof and Nyberg (2015).

[13] $\mathcal{R}$ routines are available upon request for both MAR and MARX models.

References

[1] Alessi, L., Barigozzi, M. and M. Capasso (2011), Non-Fundamentalness in Structural Econometric Models: A Review, *International Statistical Review,* 79, 1.

[2] Andrews, B., Breidt, F. and R. Davis (2006), Maximum Likelihood Estimation For All-Pass Time Series Models. *Journal of Multivariate Analysis, 97, 1638-1659.*

[3] Bork, L., Kaltwasser, P. and P. Sercu (2014), Do Exchange Rates Really Help Forecasting Commodity Prices?, *Working Paper*, available at SSRN: http://ssrn.com/abstract=2473624.

[4] Breidt, F., Davis, R., Lii, K. and M. Rosenblatt (1991), Maximum Likelihood Estimation for Noncausal Autoregressive Processes. *Journal of Multivariate Analysis, 36, 175-198.*

[5] Brockwell, P. and R. Davis (1991), *Time Series: Theory and Methods*, Springer-Verlag New York, Second Edition.

[6] Broze, L., Gouriéroux, C. and A. Szafarz (1995), Solutions of Multivariate Rational Expectation Models, *Econometric Theory*, 11, 229-257.

[7] Casella, G. and R. Berger (2002), *Statistical Inference*, Thomson Learning, Second Edition.

[8] Chen, Y., Rogoff, K. and B. Rossi (2010), Can Exchange Rates Forecast Commodity Prices?, *The Quarterly Journal of Economics*, 125(3), 1145-1194.

[9] Corradi, V. and N. Swanson (2006), Predictive Density Evaluation, *Handbook of Economic Forcasting*, Chapter 5, Vol. 1, 197-284.

[10] Cubadda, G., Hecq, A., Lieb. L. and S. Telg (2016), Serial Correlation Common Noncausal Features, *Manuscript Maastricht University.*

[11] Davis, R., Knight, K. and J. Liu (1992), M-Estimation for Autoregressions with Infinite Variance, *Stochastic Processes and Their Applications*, 40, 145-180.

[12] Davis, R. and L. Song (2012), Noncausal Vector AR Processes with Application to Economic Time Series, *Discussion Paper Colombia University.*

[13] Gouriéroux, C. and J. Jasiak (2015), Filtering, Prediction and Simulation Methods in Noncausal Processes. *Journal of Time Series Analysis*, doi: 10111/jtsa.12165.

[14] Gouriéroux, C., and J.M. Zakoïan (2016), Local Explosion Modelling by Noncausal Process, *Journal of the Royal Statistical Society, Series B*, doi:10.1111/rssb.12193.

[15] Hannan, E., Dunsmuir W., and M. Deistler (1980), Estimation of Vector ARMAX Models, *Journal of Multivariate Analysis*, 10(3), 275-295.

[16] Hansen, L. and T. Sargent (1991), Two Difficulties in Interpreting Vector Autoregressions, *Rational Expectation Econometrics*, 77-119, Westview Press, Boulder.

[17] Hecq, A., Lieb, L. and S. Telg (2016a), Identification of Mixed Causal-Noncausal Models in Finite Samples, forthcoming in *Annals of Economics and Statistics.*

[18] Hecq, A., Telg, S. and L. Lieb (2016b), Do Seasonal Adjustments Induce Noncausal Dynamics in Inflation Rates?, *MPRA Paper 74922*, University Library of Munich, Germany.

[19] Hencic, A. and C. Gouriéroux (2014), Noncausal Autoregressive Model in Application to Bitcoin/USD Exchange Rate, *Econometrics of Risk, Series: Studies in Computational Intelligence,* Springer International Publishing, 17-40.

[20] Hurvich, M. and C.L. Tsai (1989), Regression and Time Series Model Selection in Small Samples, Biometrika, 76, 297-307.

[21] Lanne, M., Luoto J. and P. Saikkonen (2012), Optimal Forecasting of Noncausal Autoregressive Time Series, *International Journal of Forecasting*, 28, 623-631.

[22] LANNE, M., NYBERG, H. AND E. SAARINEN (2012). Does Noncausality Help in Forecasting Economic Time Series?, *Economics Bulletin*, 32(4), 2849-2859.

[23] LANNE, M. AND P. SAIKKONEN (2011), Noncausal Autoregressions for Economic Time Series, *Journal of Time Series Econometrics*, 3(3), 1-32.

[24] LANNE, M. AND P. SAIKKONEN (2013), Noncausal Vector Autoregression, *Econometric Theory*, 29(3), 447-481.

[25] LOF, M. AND H. NYBERG (2015), Noncausality and the Commodity Currency Hypothesis, *Working Paper*, available at SSRN: http://ssrn.com/abstract=2597815.

[26] PESARAN, H. (2015), *Time Series and Panel Data Econometrics*, Oxford University Press.

[27] REINSEL, G. (1997), *Elements of Multivariate Time Series Analysis*, Springer Science+Business Media New York, Second Edition.

[28] WU R. AND R. DAVIS (2010), Least Absolute Deviation Estimation for General Autoregressive Moving Average Time-Series Models. *Journal of Time Series Analysis*, 31, 98-112.

Appendix

Part A - From Transfer Function Model to MARX

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

$$y_t = \psi^*(L)x_t^* + n_t, \tag{25}$$

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 $\psi^*(L) = a(L)^{-1}\theta^*(L)$. Multiplying (25) by $a(L)$

yields

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

which is the usual 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 $\varphi(z^{-1}) = -\varphi_s^* z^s \varphi^*(z)$. By rearranging terms, we find

$$
\begin{aligned}
\phi(L)\varphi(L^{-1})y_t &= \left(-\frac{1}{\varphi_s^*} + \frac{\theta_1^*}{\varphi_s^*} + \dots \frac{\theta_k^*}{\varphi_s^*}\right) x_{t+s}^* - \frac{1}{\varphi_s^*}\varepsilon_{t+s}^* \\
&= \theta(L)x_t + \varepsilon_t. \qquad (27)
\end{aligned}
$$

In case only a contemporaneous value of x_t enters the system, take $\psi^*(L) = a(L)^{-1}\beta$. Also note that the derivation can easily be extended to q regressors by defining $\psi^*(L) = [\psi_1^*(L), \dots, \psi_q^*(L)]'$ and considering X_t^*. In the distributed lag case take $\psi^*(L) = a(L)^{-1}\theta^*(L)$ with $\theta^*(L) = [\theta_1^*(L), \dots, \theta_k^*(L)]'$; in the contemporaneous case define $\psi^*(L) = a(L)^{-1}\beta$ with $\beta \in \mathbb{R}^q$. However, 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.

Part B - Proof Lemma 2: Equivalence of Information Sets for MARX Models

Proof. 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: $(i') \sim (ii')$

Note that (ii) $(y_1, ..., y_r, u_{r+1}, ..., u_T) = (y_1, ..., y_r, \phi(L)y_{r+1}, ..., \phi(L)y_T)$ by using the definition of u in equation (8). Since $u_{r+1} = y_{r+1} - \phi_1 y_r - ... - \phi_r y_1$ with $y_1, ..., y_r$ 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: $(i') \sim (iii')$

Note that (iii) $(v_1, ..., v_{T-s}, y_{T-s+1}, ..., y_T) = (\varphi(L^{-1})y_1, ..., \varphi(L^{-1})y_{T-s}, y_{T-s+1}, ..., y_T)$ by using the definition of v in equation (9). Since $v_{T-s} = y_{T-s} - \varphi_1 y_{T-s+1} - ... - \varphi_s y_T$ with $y_{T-s+1}, ..., y_T$ 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 (this reduces to the same exercise, as all information sets are augmented with the same information).

Case 3: $(ii) \sim (iv)$

Note that (iv) $(y_1, ..., y_r, \varepsilon_{r+1}, ...\varepsilon_{T-s}, u_{T-s+1}, .., u_T) =$
$(y_1, ..., y_r, \varphi(L^{-1})u_{r+1} - \beta' X_{r+1}, ..., \varphi(L^{-1})u_{T-s} - \beta' X_{T-s}, u_{T-s+1}, ..., u_T)$ by using the second equality in equation (8). Since $\varepsilon_{T-s} = u_{T-s} - \varphi_1 u_{T-s+1} - ... - \varphi_s u_T - \beta' X_{T-s}$ with $u_{T-s+1}, ..., u_T$, 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}, leading to the desired result.

Case 4: $(iii) \sim (v)$

Note that (v) $(v_1, ..., v_r, \varepsilon_{r+1}, ...\varepsilon_{T-s}, y_{T-s+1}, .., y_T) =$
$(v_1, ..., v_r, \phi(L)v_{r+1} - \beta' X_{r+1}, ..., \phi(L)v_{T-s} - \beta' X_{T-s}, y_{T-s+1}, ..., y_T)$ by using the second equality in equation (9). Since $\varepsilon_{r+1} = v_{r+1} - \phi_1 v_r - ... - \phi_r v_1 - \beta' X_{r+1}$ with $v_1, ..., v_r$, X_{r+1} and ε_{r+1} known, v_{r+1} is known. The same reasoning can be recursively applied to v_{r+2} up to v_{T-s}, leading to the desired result.

Case 5: $(i) \sim (vi)$

To show: $\tilde{y} = (y_1, ..., y_T, X_r, ..., X_{T-s}) \sim (v_1, ..., v_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, u_{T-s+1}, ..., u_T) = z$. This statement can be proven using the algebra in Lanne and Saikkonen (2012). Define the vectors $w = [v_1, ..., v_{T-s}, u_{T-s+1}, ..., u_T]'$ and $y = [y_1, ..., y_T]'$. Then

$$\begin{bmatrix} v_1 \\ \vdots \\ v_{T-s} \\ u_{T-s+1} \\ \vdots \\ v_T \end{bmatrix} = \begin{bmatrix} y_1 - \varphi_1 y_2 - ... - \varphi_s y_{s+1} \\ \vdots \\ y_{T-s} - \varphi_1 y_{T-s+1} - ... - \varphi_s y_T \\ y_{T-s+1} - \phi_1 y_{T-s} - ... - \phi_r y_{T-s+1-r} \\ \vdots \\ y_T - \phi_1 y_{T-1} - ... - \phi_r y_{T-r} \end{bmatrix} = A \begin{bmatrix} y_1 \\ \vdots \\ y_{T-s} \\ y_{T-s+1} \\ \vdots \\ y_T \end{bmatrix}, \tag{28}$$

which can be written as $w = Ay$. Now define $\tilde{x} = [\underbrace{0, ..., 0}_{r \text{ times}}, X_{r+1}, ..., X_{T-s}, \underbrace{0, ..., 0}_{s \text{ times}}]'$ Similarly, we can form the following system of equations (with slight abuse of notation, as X_t is a vector):

$$\begin{bmatrix} v_1 \\ \vdots \\ v_r \\ \varepsilon_{r+1} \\ \vdots \\ \varepsilon_{T-s} \\ u_{T-s+1} \\ \vdots \\ u_T \end{bmatrix} = \begin{bmatrix} v_1 \\ \vdots \\ v_r \\ v_{r+1} - \phi_1 v_r - ... - \phi_r v_1 - \beta' X_{r+1} \\ \vdots \\ v_{T-s} - \phi_1 v_{T-s-1} - ... - \phi_r v_{T-s-r} - \beta' X_{T-s} \\ u_{T-s+1} \\ \vdots \\ u_T \end{bmatrix} = B \begin{bmatrix} v_1 \\ \vdots \\ v_r \\ v_{r+1} \\ \vdots \\ v_{T-s} \\ u_{T-s+1} \\ \vdots \\ u_T \end{bmatrix} - \beta' \begin{bmatrix} 0 \\ \vdots \\ 0 \\ X_{r+1} \\ \vdots \\ X_{T-s} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{29}$$

or briefly $z = Bw - \beta'\tilde{x}$. Combining both systems of equations, we find that the vectors z and $\tilde{y}$ are related in the following way: $z = BAy - \beta'\tilde{x}$, where y and $\tilde{x}$ combined form the information set $\tilde{y}$. Since the matrices B and A as well as the parameter vector β only contain the known parameters, this shows that these information sets are equivalent. Combining all cases shows that information sets $(i) - (vi)$ are equivalent. □

Part C - Approximate Likelihood Function

Define $b = \beta'\tilde{x}$ such that $z = BAy - \beta'\tilde{x} = BAy - b$. Assume B and A are invertible. We are interested in the inverse transformation, i.e. $y = Q(z + b)$, where $Q = B^{-1}A^{-1}$. Let Q be a (2×2) matrix, then we have

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} q_1 & q_2 \\ q_3 & q_4 \end{bmatrix} \left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \right), \tag{30}$$

with the following functions

$$y_1 = g_1(z_1, z_2) = q_1 z_1 + q_2 z_2 + b_1 \tag{31}$$

$$y_2 = g_2(z_1, z_2) = q_3 z_1 + q_4 z_2 + b_2. \tag{32}$$

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

$$J = \begin{bmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} \\ \frac{\partial y_2}{\partial z_1} & \frac{\partial y_2}{\partial z_2} \end{bmatrix} = \begin{bmatrix} q_1 & q_2 \\ q_3 & q_4 \end{bmatrix} = Q. \tag{33}$$

Then the joint density of y_1 and y_2 is given as:

$$f_{y_1,y_2}(y_1, y_2) = \frac{1}{|\det(Q)|} f_{z_1,z_2}(BAy - b) \tag{34}$$

It is well-known that this result can be generalized to higher orders (e.g., Casella and Berger, 2002, p. 185). From Proposition 1 we know that the information sets (i) and (vi) are observationally equivalent. Using the transformations in (28)-(29) and $Q = B^{-1}A^{-1}$, we find that the probability density of the process y_t can be represented in the following way:

$$\begin{aligned} f_{y;\lambda}(y) &= \frac{1}{|\det(Q)|} f_z(BAy - b; \lambda) \\ &= |\det(A)||\det(B)| h_V(BAy - b) f_\varepsilon(BAy - b; \lambda) h_U(BAy - b), \\ &= |\det(A)| h_V(BAy - \beta'\tilde{x}) \left(\prod_{t=r+1}^{T-s} f_\sigma(BAy - \beta'\tilde{x}; \lambda) \right) h_U(BAy - \beta'\tilde{x}) \\ &= |\det(A)| h_V(\varphi(L^{-1})y_1, ..., \varphi(L^{-1})y_r) \left(\prod_{t=r+1}^{T-s} f_\sigma(\phi(L)\varphi(L^{-1})y_t - \beta' X_t; \lambda) \right) \\ &\quad h_U(\phi(L)y_{T-s+1}, ..., \phi(L)y_T), \end{aligned} \tag{35}$$

where A and B are two nonsingular matrices with $\det(B) = 1$; h_V and h_U are the joint densities of $(v_1, ..., v_r)$ and $(u_{T-s+1}, ..., u_T)$ respectively. Independence of the blocks $(v_1, ..., v_r)$, $(\varepsilon_{r+1}, ..., \varepsilon_{T-s})$ and $(u_{T-s+1}, ..., u_T)$ is applied in the second equality and the definition of the filtered values as presented in (8) and (9) in the fourth equality. Since $\det(A)$ is independent of sample size, the density of y_t can be approximated by the second term in (35).

Part D - Additional Tables and Graphs

		Specification for x_t							
		$x_t \overset{iid}{\sim} t(5)$		$x_t \overset{iid}{\sim} N(0,1)$		$x_t \overset{iid}{\sim} C(0,1)$		$x_t \sim$ AR(1)	
T	Parameter	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
50	ϕ	0.313	0.168	0.309	0.168	0.297	0.073	0.291	0.078
	φ	0.461	0.164	0.469	0.156	0.491	0.072	0.495	0.065
	β	0.305	0.158	0.306	0.187	0.301	0.037	0.304	0.037
	ν	5.022	7.790	5.188	8.935	5.158	7.936	5.750	11.368
100	ϕ	0.302	0.110	0.301	0.113	0.300	0.039	0.296	0.052
	φ	0.489	0.103	0.484	0.106	0.497	0.036	0.499	0.042
	β	0.301	0.104	0.306	0.135	0.300	0.018	0.300	0.024
	ν	3.477	1.460	3.388	1.463	3.494	1.594	3.659	3.072
500	ϕ	0.301	0.037	0.300	0.040	0.300	0.016	0.299	0.022
	φ	0.497	0.034	0.498	0.036	0.500	0.008	0.500	0.017
	β	0.300	0.043	0.302	0.056	0.300	0.004	0.300	0.010
	ν	3.053	0.352	3.070	0.386	3.069	0.413	3.056	0.382
1000	ϕ	0.300	0.026	0.300	0.025	0.300	0.005	0.299	0.015
	φ	0.499	0.023	0.500	0.024	0.500	0.004	0.500	0.013
	β	0.300	0.030	0.300	0.038	0.300	0.002	0.300	0.007
	ν	3.039	0.254	3.033	0.244	3.027	0.281	3.031	0.258

Table 6: Finite sample properties of the ML estimator for an MARX(1,1,1) with $\varepsilon_t \overset{iid}{\sim} t(3)$

Part E - Proofs of Lemma 2 and Theorem 1

Proof of Lemma 2

For the proofs, we need some additional notation. Define $e_t = \frac{f'_\sigma(\varepsilon_t;\lambda)}{f_\sigma(\varepsilon_t;\lambda)} = \frac{f'(\varepsilon_t/\sigma;\lambda)}{\sigma f(\varepsilon_t/\sigma;\lambda)}$, $\tilde{\mathcal{J}} = \sigma^{-2}\mathcal{J}$ and $\tilde{\mathcal{I}} = \sigma^{-2}\mathcal{I}$. Furthermore, let $x = \varepsilon_t/\sigma$, then we have that

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

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

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

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

$$\mathbb{E}(z_s e_t) = \begin{cases} 0, & \text{if } s \neq t, \\ -1, & \text{if } s \neq t, \end{cases} \tag{36}$$

which follows from the assumptions on the density and strixt exogeneity between all exogeneous regressors and the error term. Now, for $i = 1, ..., r$, we can show that

$$\mathbb{E}\left(\frac{\partial g_t(\theta_0)}{\partial \phi_i}\right) = \mathbb{E}(-e_t v_{t-i})$$

$$= -\mathbb{E}\left(e_t \sum_{j=0}^{\infty} \alpha_j z_{t-i-j}\right)$$

$$= 0.$$

Hence, we note that V_{t-1} and e_t are still independent as in Lanne and Saikkonen (2011). Consequently, we still find

$$\begin{aligned}
\text{Cov}\left(\frac{\partial g_t(\theta_0)}{\partial \phi}\right) &= \text{Cov}(-V_{t-1}e_t) \\
&= \mathbb{E}(e_t^2)\mathbb{E}(V_{t-1}V_{t-1}') \\
&= \tilde{\mathcal{J}}\Gamma_V,
\end{aligned}$$

where Γ_V denotes the autocovariance matrix of the vector V_{t-1}. Because $V_{t-1}e_t$ is uncorrelated, we have

$$\lim_{T\to\infty}(T-p)^{-1}\text{Cov}\left(\sum_{t=r+1}^{T-s} \frac{\partial g_t(\theta_0)}{\partial \phi}\right) = \tilde{\mathcal{J}}\Gamma_V.$$

Symmetrically, by using similar arguments, we can show for $i = 1, ..., s$ that

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial g_t(\theta_0)}{\partial \varphi_i}\right) &= \mathbb{E}(-e_t u_{t+i}) \\
&= -\mathbb{E}\left(e_t \sum_{j=0}^{\infty} \delta_j z_{t+i+j}\right) \\
&= 0.
\end{aligned}$$

That is, the independence of e_t and U_{t+1} is preserved through strict exogeneity. Letting Γ_U be the autocovariance matrix of U_{t+1}, we find that

$$\text{Cov}\left(\frac{\partial g_t(\theta_0)}{\partial \varphi}\right) = \text{Cov}(-U_{t+1}e_t)$$

$$= \mathbb{E}(e_t^2)\mathbb{E}(U_{t+1}U_{t+1}')$$

$$= \tilde{\mathcal{J}}\Gamma_U.$$

Because $U_{t+1}e_t$ is uncorrelated, we have

$$\lim_{T\to\infty}(T-p)^{-1}\text{Cov}\left(\sum_{t=r+1}^{T-s}\frac{\partial g_t(\theta_0)}{\partial\varphi}\right) = \tilde{\mathcal{J}}\Gamma_U.$$

Lastly, we can apply the same logic for the parameter vector β. Since for $i = 1, ..., q$, we have that

$$\mathbb{E}\left(\frac{\partial g_t(\theta_0)}{\partial\beta_i}\right) = 0$$

by the independence of $x_{i,t}$ and ε_t. If we denote by Γ_X, the autocovariance matrix of X_t, it follows that

$$\text{Cov}\left(\frac{\partial g_t(\theta_0)}{\partial\beta}\right) = \text{Cov}(-X_te_t)$$

$$= \mathbb{E}(e_t^2)\mathbb{E}(X_tX_t')$$

$$= \tilde{\mathcal{J}}\Gamma_X.$$

Because X_te_t is uncorrelated, we have

$$\lim_{T\to\infty}(T-p)^{-1}\text{Cov}\left(\sum_{t=r+1}^{T-s}\frac{\partial g_t(\theta_0)}{\partial\beta}\right) = \tilde{\mathcal{J}}\Gamma_X.$$

We now characterize the covariances of the partials. To that end, we first notice that

$$\mathrm{Cov}(z_{t-i}e_t, z_{k-j}e_k) = \begin{cases} \mathcal{I} + \sum_{m=1}^{q}\sum_{l=1}^{q}\beta_m\beta_l\gamma_{x_m,x_l}(0)\tilde{\mathcal{J}} & \text{if } t=k, i=j=0, \\ \mathcal{J} + \sum_{m=1}^{q}\sum_{l=1}^{q}\beta_m\beta_l\gamma_{x_m,x_l}(0)\tilde{\mathcal{J}} & \text{if } t=k, i=j\neq 0, \\ 1 & \text{if } t\neq k, i=t-k, j=k-t, \\ 0 & \text{otherwise.} \end{cases} \tag{37}$$

Hence, using (36)-(37), we find that

$$\mathrm{Cov}\left(\frac{\partial g_t(\theta_0)}{\partial \phi_i}, \frac{\partial g_k(\theta_0)}{\partial \phi_j}\right) = \begin{cases} \gamma_V(i-j)\tilde{\mathcal{J}}, & \text{if } k=t, 1\le i\le j\le r, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathrm{Cov}\left(\frac{\partial g_t(\theta_0)}{\partial \varphi_i}, \frac{\partial g_k(\theta_0)}{\partial \varphi_j}\right) = \begin{cases} \gamma_U(i-j)\tilde{\mathcal{J}}, & \text{if } k=t, 1\le i\le j\le s, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathrm{Cov}\left(\frac{\partial g_t(\theta_0)}{\partial \beta_i}, \frac{\partial g_k(\theta_0)}{\partial \beta_j}\right) = \begin{cases} \gamma_{x_i,x_j}(0)\tilde{\mathcal{J}}, & \text{if } k=t, 1\le i\le j\le q, \\ 0, & \text{otherwise.} \end{cases}$$

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

$$\begin{aligned}\mathrm{Cov}\left(\frac{\partial g_t(\theta_0)}{\partial \phi_i}, \frac{\partial g_k(\theta_0)}{\partial \varphi_j}\right) &= \mathrm{Cov}\left(\sum_{a=0}^{\infty}\alpha_a z_{t-i-a}e_t, \sum_{b=0}^{\infty}\delta_b z_{k+j+b}e_k\right) \\ &= \sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\alpha_a\delta_b \mathrm{Cov}\left(z_{t-i-a}e_t, z_{k+j+b}e_k\right)\end{aligned}$$

$$= \begin{cases} \alpha_{t-k-i}\delta_{t-k-j}, & \text{for } t>k, 1\le i\le r, 1\le j\le s. \\ \tilde{\mathcal{J}}\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\alpha_a\delta_b\sum_{m=1}^{q}\sum_{l=1}^{q}\beta_m\beta_l\gamma_{x_m,x_l}(t-i-a, t+j+b) & \text{for } t=k, 1\le i\le r, 1\le j\le s. \\ 0 & \text{for } t<k, 1\le i\le r, 1\le j\le s. \end{cases}$$

The element (i,j) of the matrix $(T-p)^{-1}\mathrm{Cov}(\partial L_T(\theta_0)/\partial\phi, \partial L_T(\theta_0)/\partial\varphi)$ is

$$
\begin{aligned}
&(T-p)\mathrm{Cov}\left((T-p)^{-1}\sum_{t=r+1}^{T-s}\frac{\partial g_t(\theta_0)}{\partial\phi_i}, (T-p)^{-1}\sum_{k=r+1}^{T-s}\frac{\partial g_k(\theta_0)}{\partial\varphi_j}\right)\\
&=(T-p)^{-1}\sum_{t=r+1}^{T-s}\sum_{k=r+1}^{T-s}\mathrm{Cov}\left(\frac{\partial g_t(\theta_0)}{\partial\phi_i}, \frac{\partial g_k(\theta_0)}{\partial\varphi_j}\right)\\
&=(T-p)^{-1}\sum_{t=r+1}^{T-s}\sum_{k=r+1}^{T-s}\mathrm{Cov}(-v_{t-i}e_t, -u_{k+j}e_k)\\
&=(T-p)^{-1}\sum_{t=r+1}^{T-s}\sum_{k=r+1}^{T-s}\left(\mathbb{1}_{\{t>k\}}\alpha_{t-k-i}\delta_{t-k-j}+\mathbb{1}_{\{t=k\}}\tilde{\mathcal{J}}\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\alpha_a\delta_b\sum_{m=1}^{q}\sum_{l=1}^{q}\beta_m\beta_l\gamma_{x_m,x_l}(t-i-a,k+j+b)\right)\\
&=(T-p)^{-1}\left(\sum_{k=r+1}^{T-s-1}\sum_{t=k+1}^{T-s}\alpha_{t-k-i}\delta_{t-k-j}+\sum_{t'=r+1}^{T-s}\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\alpha_a\delta_b\tilde{\mathcal{J}}\sum_{m=1}^{q}\sum_{l=1}^{q}\beta_m\beta_l\gamma_{x_m,x_l}(t'-i-a,t'+j+b)\right)\\
&=(T-p)^{-1}\left(\sum_{k=r+1}^{T-s-1}\sum_{t=0}^{T-s-k-i}\alpha_t\delta_{t+i-j}+\tilde{\mathcal{J}}\sum_{t'=r+1}^{T-s}\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\alpha_a\delta_b\sum_{m=1}^{q}\sum_{l=1}^{q}\beta_m\beta_l\gamma_{x_m,x_l}(t'-i-a,t'+j+b)\right)\\
&\to\sum_{t=0}^{\infty}\alpha_t\delta_{t+i-j}+(T-p)^{-1}\tilde{\mathcal{J}}\sum_{t'=r+1}^{T-s}\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\alpha_a\delta_b\sum_{m=1}^{q}\sum_{l=1}^{q}\beta_m\beta_l\gamma_{x_m,x_l}(t'-i-a,t'+j+b),
\end{aligned}
$$

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 $t>k$. If we assume that the q processes in X_t are jointly covariance stationary, the cross covariance and cross correlation is independent of time. This reduces the above expression to the following:

$$\to\sum_{t=0}^{\infty}\alpha_t\delta_{t+i-j}+\tilde{\mathcal{J}}\sum_{a=0}^{\infty}\sum_{b=0}^{\infty}\alpha_a\delta_b\sum_{m=1}^{q}\sum_{l=1}^{q}\beta_m\beta_l\gamma_{x_m,x_l}(i+j+a+b).$$

That is, the sum over time can now be replaced by $(T-p)$ as the cross covariance does not depend on time anymore. This $(T-p)$ cancels with the $(T-p)^{-1}$ in front of the whole expression. Because of geometric decay of the sequences $\{\alpha_a\}$ and $\{\delta_b\}$ we know that the sequence converges.

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,

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

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:

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

To prove asymptotic normality, first define $M = diag(\Sigma, \Omega)$. By the Cramér-Wold theorem, it suffices to show that for any vector a of appropriate size,

$$\frac{1}{\sqrt{T-p}} \sum_{t=r+1}^{T-s} a' \frac{\partial g_t(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, a'Ma). \tag{38}$$

Define the sequence of $(p+q+d+1)$ dimensional random vectors $\{W_{tm}, t \in \mathbb{Z}\}$ to be the partials defined in Section 3.1, where v_t and u_t are replaced by their representation in (11)-(12) with the sums truncated at a large positive integer m, i.e.,

$$v_t = \sum_{j=0}^{m} \alpha_j z_{t-j} \quad \text{and} \quad u_t = \sum_{j=0}^{m} \delta_j z_{t+j},$$

where $z_t = \varepsilon_t + \sum_{i=1}^{q} \beta_i x_{i,t}$. Let M_m be the matrix corresponding to M, obtained by truncating u_t and v_t. Then the stationary sequence $\{W_{tm}, t \in \mathbb{Z}\}$ is $\max\{2m+p, \ell\}$ dependent[14]. Applying Theorem 6.4.2 in Brockwell and Davis (1991), we get

$$\frac{1}{\sqrt{T-p}} \sum_{t=r+1}^{T-s} a' W_{tm} \xrightarrow{d} N(0, a' M_m a).$$

Now, since $M_m \to M$ as $m \to \infty$ and since

$$\lim_{m\to\infty} \lim_{T\to\infty} \text{Var}\left(\frac{1}{\sqrt{T-p}} \sum_{t=r+1}^{T-s} \left(a' W_{tm} - a' \frac{\partial g_t(\theta_0)}{\partial \theta} \right) \right) = 0,$$

the convergence in (38) is immediate from Proposition 6.3.9 in Brockwell and Davis (1991).

Proof of Theorem 1

Similar to Lanne and Saikkonen (2011), we first present the second partial derivatives of the function $g_t(\theta)$. We 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 Y_t be the $(r \times s)$ matrix with elements y_{t-i+j}. Write $\tilde{v}_t = v_t(\varphi)$ and $\tilde{u}_t = u_t(\phi)$ and thus $\tilde{V}_{t-1} = (\tilde{v}_{t-1}, ..., \tilde{v}_{t-r})$ and $\tilde{U}_{t+1} = (\tilde{u}_{t+1}, ..., \tilde{u}_{t+s})$ to simplify notation. Similarly, $\tilde{\varepsilon}_t = \tilde{v}_t - \phi_1 \tilde{v}_{t-1} - ... - \phi_r \tilde{v}_{t-r} = \tilde{u}_t - \varphi_1 \tilde{u}_{t+1} - ... - \varphi_s \tilde{u}_{t+s}$ denotes ε_t evaluated at an arbitrary point in the permissible parameter space, not

[14]The $2m+p$ follows from writing u_t and v_t in their truncated representation, ℓ follows from the processes in X_t which are assumed to be at most ℓ-dependent sequences

the true one. Then, the second partial derivatives in the MARX case can be obtained through straightforward differentiation, similar Lanne and Saikkonen (2011) and Breidt et al. (1991):

$$\partial^2 g_t(\theta)/\partial\phi\partial\phi' = \sigma^{-2}h'(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{V}_{t-1}\tilde{V}_{t-1}'$$

$$\partial^2 g_t(\theta)/\partial\varphi\partial\varphi' = \sigma^{-2}h'(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{U}_{t+1}\tilde{U}_{t+1}'$$

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

$$\partial^2 g_t(\theta)/\partial\sigma^2 = 2\sigma^{-3}h(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{\varepsilon}_t + \sigma^{-4}h'(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{\varepsilon}_t^2 + \sigma^{-2}$$

$$\partial^2 g_t(\theta)/\partial\lambda\partial\lambda' = \frac{1}{f(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)}\frac{\partial^2 f(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)}{\partial\lambda\partial\lambda'} - \frac{1}{f^2(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)}\left(\frac{\partial f(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)}{\partial\lambda}\right)\left(\frac{\partial f(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)}{\partial\lambda}\right)'$$

$$\partial^2 g_t(\theta)/\partial\phi\partial\varphi' = \sigma^{-2}h'(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{V}_{t-1}\tilde{U}_{t+1}' + \sigma^{-1}h(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)Y_t$$

$$\partial^2 g_t(\theta)/\partial\phi\partial\beta' = \sigma^{-2}h'(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{V}_{t-1}X_t'$$

$$\partial^2 g_t(\theta)/\partial\phi\partial\sigma = \sigma^{-3}h'(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{\varepsilon}_t\tilde{V}_{t-1} + \sigma^{-2}h(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{V}_{t-1}$$

$$\partial^2 g_t(\theta)/\partial\phi\partial\lambda' = -\sigma^{-1}\tilde{V}_{t-1}\partial h(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)/\partial\lambda'$$

$$\partial^2 g_t(\theta)/\partial\varphi\partial\beta' = \sigma^{-2}h'(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{U}_{t+1}X_t'$$

$$\partial^2 g_t(\theta)/\partial\varphi\partial\sigma = \sigma^{-3}h'(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{\varepsilon}_t\tilde{U}_{t+1} + \sigma^{-2}h(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{U}_{t+1}$$

$$\partial^2 g_t(\theta)/\partial\varphi\partial\lambda' = -\sigma^{-1}\tilde{U}_{t+1}\partial h(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)/\partial\lambda'$$

$$\partial^2 g_t(\theta)/\partial\beta\partial\sigma = \sigma^{-3}h'(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)\tilde{\varepsilon}_tX_t + \sigma^{-2}h(\sigma^{-1}\tilde{\varepsilon}_t;\lambda)X_t$$

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

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

It can be verified that $\mathbb{E}(\partial^2 g_t(\theta_0)/\partial\theta\partial\theta') = -diag(\Sigma,\Omega)$. 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

$$\sum_{t=r+1}^{T-s}\Big[g_t(\theta_0 + T^{-1/2}c) - g_t(\theta_0)\Big] = T^{-1/2}\sum_{t=r+1}^{T-s} c'\frac{\partial g_t(\theta_0)}{\partial\theta} + \frac{1}{2}T^{-1}\sum_{t=r+1}^{T-s} c'\frac{\partial^2 g_t(\theta_0)}{\partial\theta\partial\theta'}c$$

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

where $c \in \mathbb{R}^{r+s+q+1+d}$ and the argument $\theta_T^*(c)$ in the matrix of second partial derivatives means that each row is evaluated at an intermediate point lying between the true parameter value θ_0 and $T^{-1/2}c$. If $\|\cdot\|$ denotes the Euclidian norm we have $\sup_{c\in C}\|\theta_T^*(c) - \theta_0\| \to 0$ for any compact set $C \subset \mathbb{R}^{r+s+q+1+d}$. Using the dominance conditions (A7) in Davis et al. (2006), arguments similar to Breidt et al. (1991, p. 186-190) and assumption (A1) in this paper, it can be shown that a uniform law of large numbers for stationary ergodic processes applies to $\partial^2 g_t(\theta)/\partial\theta\partial\theta'$ over any small enough compact neighborhood θ_0. We can conclude that

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

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