Mixed Causal-Noncausal Autoregressions with Strictly Exogenous Regressors

Hecq, Alain and Issler, João Victor and Telg, Sean

Maastricht University, Graduate School of Economics - EPGE, Maastricht University

11 August 2017

Mixed Causal-Noncausal Autoregressions with Strictly Exogenous Regressors

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

August 11, 2017

Abstract

The mixed autoregressive causal-noncausal model (MAR) has been proposed to estimate economic relationships involving explosive roots in their autoregressive part, as they have stationary forward solutions. In previous work, possible exogenous variables in economic relationships are substituted into the error term to ensure the univariate MAR structure of the variable of interest. To allow for the impact of exogenous fundamental 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 influence of the U.S. exchange rate and the U.S. industrial production index on several commodity prices.

Keywords: Mixed causal-noncausal process, non-Gaussian errors, identification, rational expectation models, 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, 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.

Part of this work has been written while Sean Telg was visiting the CREST in Paris and Alain Hecq the EPGE/FGV in Rio de Janeiro. We thank both 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) for valuable comments and remarks.

1 Introduction

The usefulness of mixed causal-noncausal autoregressive (MAR) models in time series econometrics can be explained by three reasons. First, Gouriéroux and Zakoïan (2016), Hencic and Gouriéroux (2014) and Hecq et al. (2016a) demonstrate how the inclusion of noncausal autoregressive terms can generate dynamic patterns like speculative bubbles and asymmetric cycles that instead should be generated using complex nonlinear models. Second, Lanne et al. (2012a; 2012b) 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 causal autoregressive models take only the fundamental solution of a system into account, the mixed causal-noncausal autoregressive 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 the one of econometricians.

The inclusion of exogenous regressors might further improve the relevance of the MAR model for modelling and forecasting economic processes. This model, enriched with strictly exogenous variables, is denoted MARX. In this paper, we show how the parameters of the MARX model can be estimated by maximum likelihood (ML) for noise driven by a general class of non-Gaussian densities as defined in Andrews et al. (2006). For the particular case of a Student's t-likelihood, we provide a method to compute closed form solutions of the corresponding standard errors. This is also done for the Least Absolute Deviation (LAD) estimator, as it is often used as an initial estimator (Lanne and Saikkonen, 2011) and is known to outperform the Student's t ML estimator in certain instances (Hecq et al., 2016a). We show that purely causal and noncausal models with additional regressors have the appealing feature to be (potentially) identifiable under

Gaussianity, due to cross-covariances creating an asymmetry in the autocovariance function of both models. This property is not shared by purely causal and noncausal models without exogenous regressors. By means of Monte Carlo simulations, we evaluate the performance of the ML estimator as well as our proposed model selection method based on information criteria. The methods proposed in this paper are implemented in the R package **MARX**; its specifications are discussed in detail in Hecq et al. (2017).[1]

A simple underlying economic example where the MARX model can be used 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). More generally, many relationships that involve expectation terms and consequently have a present value solution, fall in this framework. In the empirical section we investigate the relationship between several commodity prices indices representative of the global market ($cp_t \equiv \ln CP_t$) and two fundamental explanatory variables: the U.S. nominal exchange rate ($s_t \equiv \ln S_t$) and the U.S. industrial production index ($ip_t \equiv \ln IPI_t$) (see Chen et al., 2010; Bork et al., 2014).[2] Let us consider a rational expectation (RE) model such as

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

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.[3] Using the quadratic determinantal equation method, we can write (1) as

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

[1] The package is freely available at `https://CRAN.R-project.org/package=MARX`.

[2] 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 U.S. exchange rate is released by the Federal Reserve Bank of St. Louis. We also make use of the industrial production index from the FRED database.

[3] We can substitute the expectation term $\mathbb{E}(\Delta cp_{t+1}|\Omega_t)$ either by a perfect foresight scheme $\mathbb{E}(\Delta cp_{t+1}|\Omega_t) = \Delta cp_{t+1}$ or by the 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 et al., 1995).

where α_b is one of the two roots of the quadratic equation $\beta_f x^2 - x + \beta_b = 0$ (for more details see Pesaran, 2015, p. 473-475). That is, the RE model can be represented as a lag-augmented present value model where commodity prices Granger-cause exchange rates: a model compatible with our MARX representation.

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. Section 6 summarizes and concludes. Proofs and additional material are collected 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 $x_{i,t}$ $(i = 1, ..., q)$ be the ith variable in a set of q for y_t and $\boldsymbol{\beta} \in \mathbb{R}^q$ a vector of parameters. Then we can define $\boldsymbol{X}_t = [x_{1,t}, ..., x_{q,t}]' \in \mathbb{R}^q$ as the vector of all exogenous variables at time t.[4] The MARX(r, s, q) for a stationary time series y_t can be represented as

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

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

[4] We only consider contemporaneous values of $\boldsymbol{X}_t$ 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.

model with strictly exogenous regressors, denoted MARX$(r, 0, q)$ or simply ARX(r, q):

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

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

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

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 the general 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 $\boldsymbol{X}_t$ and ε_t. That is,

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

where $z_{t-j} = \varepsilon_{t-j} + \sum_{i=1}^{q} \beta_i x_{i,t-j}$ and $\pi(z, z^{-1})$ is a polynomial satisfying $\pi(z, z^{-1})\phi(z)\varphi(z^{-1}) = 1$. Note that we replace the operator L by the complex variable z when considering the properties of polynomials. Similar to Gouriéroux and Jasiak (2015), we note that the polynomials $\phi(z)$ and $\varphi(z^{-1})$ are invertible and their inverses create infinite series in z and z^{-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 $\boldsymbol{X}_t$. 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 (7) consists 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 $[\phi(z)\varphi(z^{-1})]$ to $\pi(z, z^{-1})$.[5]

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:*

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

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

In order to ensure identifiability of the parameter vector $\boldsymbol{\beta}$ and to prove consistency of the ML estimator, we make the following assumptions on $\boldsymbol{X}_t$:

Assumptions. The processes in $\boldsymbol{X}_t$ are assumed to be

(A1) ergodic, (strictly) stationary and strictly exogenous w.r.t. ε_t ($\boldsymbol{X}_t$ and ε_t are independent stochastic processes);

(A2) mixed causal-noncausal with finite variance, i.e., $x_{i,t} = c_i + \sum_{j=-\infty}^{\infty} \rho_{i,j}\eta_{i,t-j}$ with $\eta_{i,t}$ strong white noise;

(A3) linearly independent.

Assumption (A1) is necessary for the Central Limit Theorem for m-dependent processes (Theorem 6.4.2 in Brockwell and Davis, 1991). Assumption (A2) defines the dynamic structure the processes in $\boldsymbol{X}_t$ can assume, allowing for a two-sided moving process, purely causal and

[5]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 et al. (2016b).

noncausal as well as mixed ARMA processes are allowed. By assumption (A3) these processes are linearly independent, which simplifies the proof for consistency of the ML estimator. This assumption can be relaxed, but we leave this for future research, as it is not restrictive for the empirical application considered in this paper.

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 $\boldsymbol{X}_t$ 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 values. 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 $\phi(z)\varphi(z^{-1})$. 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 (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+1} and $x^*_{1,T}$.[6] In the general MARX(r, s, q) setup, filtered values are used to circumvent the problem. Defining $[\varphi(z^{-1})]^{-1} \equiv \delta(z^{-1})$, we can rewrite the second equality in (8) in the following way:

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

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

In a similar fashion, when we take $[\phi(z)]^{-1} \equiv \alpha(z)$, 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, v) can be simulated independently and can be interpreted as a causal [noncausal] "error term" of a purely noncausal [causal] autoregression.

We characterize the simulation steps using (8). Note that the first equality $\phi(L)y_t = u_t$ appears like a conventional causal autoregressive process. In order to simulate such a process, one needs r starting values, say $y^*_{-1}, ..., y^*_{-r}$, to create the first value y_1. Additionally, one needs the value v_1 which is usually a draw from a desired distribution in the case of a causal autoregressive process. In the MARX case, however, v_1 is represented as a linear combination of current and future values of ε_t and $\boldsymbol{X}_t$ as can be seen in (11). If we consider a truncation m sufficiently large for the infinite sum, v_1 can be constructed by simulating long paths of ε_t and $\boldsymbol{X}_t$ such that z_1 up to z_{1+m} are available. Now that y_1 is generated, it can be used to construct the next value y_2. In general, the following two steps can be used to simulate an MARX processes based on (8):

1. Generate a path of length $(T + m)$ for ε_t and $\boldsymbol{X}_t$ and simulate the values of u_t using a truncated version of (11).

2. Create *starting* values y^*_{-1} up to y^*_{-r} and simulate the process y_t like a conventional causal autoregressive process.

It is also possible to base the simulation on (9), then one typically generates the series "backwards". In that case, the following steps are needed:

1. Generate a path of length $(T + m)$ for ε_t and $\boldsymbol{X}_t$ and simulate the values of v_t using a

truncated version of (12).

2. Create *terminal* values y^*_{T+1} up to y^*_{T+s} and simulate the process y_t like a noncausal autoregressive process (see e.g., Gouriéroux and Jasiak, 2015).

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

$$y_t = \varphi_1 y_{t+1} + \sum_{j=0}^{\infty} \phi^j (\beta_1 x_{1,t-j} + \varepsilon_{t-j}), \tag{13}$$

where a truncation m sufficiently large has to be considered for the infinite sum.[7] *The simulation of this process consists of two steps. Firstly, simulate a long path of length $T+m$ for $x_{1,t}$ and ε_t. Accordingly, one can construct the sequence v_t for $t = 1, ..., T$. Secondly, choose a terminal value, say y^*_{T+1}. Using* (13)*, we are going to simulate the series "backwards". That is, to simulate y_T, we put the terminal value y^*_{T+1} for y_{t+1} in* (13) *and use the value v_T simulated in the first step. Now that y_T becomes available, it can be used in combination with v_{T-1} to construct y_{T-1}. We continue this procedure until we reach y_1.*

Figure 1 shows simulated paths of an MARX(1,1,1) and MAR(1,1) process with $[\phi_1, \varphi_1]' = [0.3, 0.9]'$, $\beta_1 = 0.3$, $x_{1,t} \overset{iid}{\sim} t(1,0)$ and $\varepsilon_t \overset{iid}{\sim} t(3,0)$ with $t(\nu, \sigma)$ being the Student's t distribution with degrees of freedom parameter ν and scale parameter σ. The truncation m is set at 10,000. 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 $x_{1,t}$, 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 shows

[7] Since $\deg(\phi(z)) = 1$ 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.

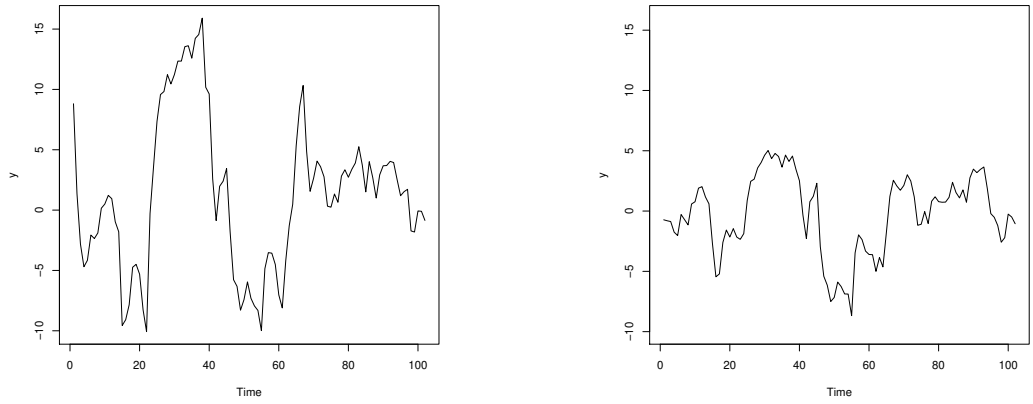

Figure 1: A simulated MARX process (left) and the same process without exogenous variable (right)

the equivalence of various information sets. A proof of this proposition is in the Appendix C.

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

(i) $(y_1, ..., y_T, \boldsymbol{X}_{r+1}, ..., \boldsymbol{X}_{T-s})$

(ii) $(y_1, ..., y_r, u_{r+1}, ..., u_T, \boldsymbol{X}_{r+1}, ..., \boldsymbol{X}_{T-s})$

(iii) $(v_1, ..., v_{T-s}, y_{T-s+1}, ..., y_T, \boldsymbol{X}_{r+1}, ..., \boldsymbol{X}_{T-s})$

(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 and spectral density (which are 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, an interesting property of pure MARX models is the possibility to identify models 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 k-th order autocovariance for the causal model, denoted by superscript C, can be written as

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

and for the noncausal model, depicted as superscript NC, as

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

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 that identification cannot be achieved. The cross-covariances $\gamma_{xy}(k)$ and $\gamma_{xy}(-k)$ however need not be equal (and rarely are equal), which causes them to create a different autocovariance structure for a causal and noncausal data generating process.

To construct the autocovariance for the purely causal MARX(1,0,1), we multiply the model by y_{t-k} and take expectations. Since y_{t-k} does not depend on ε_t, the expectation of their product equals zero. In the purely noncausal case, similar logic holds when the model is multiplied by y_{t+k}. Hence, we exploit in both cases the existence of a y_s, which is independent of ε_t, $s \neq t$. Since in the mixed causal-noncausal case y_t depends on its past, current and future errors, no such argument can be used.

3 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).[8] 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 assume that the density function $f(x;\boldsymbol{\lambda})$ satisfies the regularity conditions of Andrews et al. (2006).[9] The permissible parameter space of $\boldsymbol{\lambda}$,

[8] 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 $\boldsymbol{X}_t$.

[9] 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.

denoted by $\boldsymbol{\Lambda}$, is some subset of $\mathbb{R}^d$. The scale parameter 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 (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 Appendix B 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:

$$\begin{aligned} L_T(\theta) &= \sum_{t=r+1}^{T-s} \ln f_\sigma(\phi(L)\varphi(L^{-1})y_t - \boldsymbol{\beta}'\boldsymbol{X}_t; \boldsymbol{\lambda}) \\ &= \sum_{t=r+1}^{T-s} g_t(\boldsymbol{\theta}), \end{aligned} \tag{14}$$

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 as in (8) and (9) 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:

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

where we also used $f_\sigma(x; \boldsymbol{\lambda}) = \sigma^{-1} f(\sigma^{-1}x; \boldsymbol{\lambda})$ (definition of density). 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 $\boldsymbol{\Lambda}$.

3.1 Asymptotic Properties of the AML Estimator

We first consider the score of $\boldsymbol{\theta}$ evaluated at true parameter values. Define $\boldsymbol{V}_{t-1} = [v_{t-1}, ..., v_{t-r}]'$ and $\boldsymbol{U}_{t+1} = [u_{t+1}, ..., u_{t+s}]'$, 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 (14), we obtain:

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

and

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

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

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

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

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

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 $\boldsymbol{V}_{t-1}$ and $\boldsymbol{U}_{t+1}$

respectively. $\mathbf{\Gamma}_X$ is the cross-covariance matrix of the q processes in $\boldsymbol{X}_t$ which is diagonal under the assumption of linear independence between processes in $\boldsymbol{X}_t$. $\mathbf{\Sigma}_{12}$ 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 $\mathbf{\Sigma}_{13}$ matrix has size $(r \times q)$ with the (i,j)th element equal to

$$\beta_j \sigma^{-2} \mathcal{J} \sum_{a=0}^{\infty} \alpha_a \gamma_{x_j}(i+a),$$

while for $\mathbf{\Sigma}_{23}$ this element is

$$\beta_j \sigma^{-2} \mathcal{J} \sum_{b=0}^{\infty} \delta_b \gamma_{x_j}(i+b),$$

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

$$\mathbf{\Omega} = \begin{bmatrix} \omega_\sigma^2 & \boldsymbol{\omega}_{\sigma\lambda} \\ \boldsymbol{\omega}_{\lambda\sigma} & \mathbf{\Omega}_{\lambda\lambda} \end{bmatrix},$$

where

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

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

and

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

Lemma 2. *If conditions (A1)-(A7) of Andrews et al. (2006) and assumptions (A1)-(A3) hold,*

then

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

Moreover, the matrices $\boldsymbol{\Sigma}$ *and* $\boldsymbol{\Omega}$ *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 $\mathcal{J} > 1$.

Theorem 1. *If conditions (A1)-(A7) of Andrews et al. (2006) and assumptions (A1)-(A3) hold, there exists a sequence of local maximizers* $\hat{\boldsymbol{\theta}} = [\hat{\boldsymbol{\phi}}', \hat{\boldsymbol{\varphi}}', \hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\lambda}}', \hat{\sigma}]'$ *of* $L_T(\boldsymbol{\theta})$ *in* (14) *such that*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta_0}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, diag(\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Omega}^{-1})).$$

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 $[\boldsymbol{\phi}', \boldsymbol{\varphi}', \boldsymbol{\beta}']'$ and the distributional parameters $[\sigma, \boldsymbol{\lambda}']'$ 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 parameters of both the autoregressive and exogenous variables for Student's t-MLE and the LAD estimator.

3.2.1 Student's t Maximum Likelihood Estimation

Similar to Hecq et al. (2016a), 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 $(n \times 1)$ series $\boldsymbol{u} \equiv \boldsymbol{U}_t^* = [u_{r+1}, ..., u_{T-s}]'$ up to $\boldsymbol{U}_{t+s}^* = [u_{r+s+1}, ..., u_T]'$, $\boldsymbol{V}_{t-r}^* = [v_1, ..., v_{T-p}]'$ up to $\boldsymbol{v} \equiv \boldsymbol{V}_t^* = [v_{r+1}, ..., v_{T-s}]'$, $\boldsymbol{X}_{i,t} = [x_{i,r+1}, ..., x_{i,T-s}]'$ and $\boldsymbol{\varepsilon} = [\varepsilon_{r+1}, ..., \varepsilon_{T-s}]'$. We construct $\boldsymbol{Z} = [\boldsymbol{U}_{t+1}^*, ..., \boldsymbol{U}_{t+s}^*, \boldsymbol{X}_{1,t}, ..., \boldsymbol{X}_{q,t}]$ and similarly $\boldsymbol{Q} = [\boldsymbol{V}_{t-1}^*, ..., \boldsymbol{V}_{t-r}^*, \boldsymbol{X}_{1,t}, ..., \boldsymbol{X}_{q,t}]$, which are of dimensions $(n \times (s+q))$ and $(n \times (r+q))$ respectively. Using this notation, we can write the autoregressions defined in (8) and (9) in matrix notation as follows:

$$\boldsymbol{u} = \boldsymbol{Z}\boldsymbol{\zeta} + \boldsymbol{\varepsilon}, \tag{15}$$

$$\boldsymbol{v} = \boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{16}$$

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

$$\sqrt{n}(\hat{\boldsymbol{\zeta}}_{ML} - \boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \frac{\nu+3}{\nu+1}\sigma^2 \boldsymbol{\Upsilon}_{\boldsymbol{\phi}}^{-1}\right), \tag{17}$$

$$\sqrt{n}(\hat{\boldsymbol{\xi}}_{ML} - \boldsymbol{\xi}_0) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \frac{\nu+3}{\nu+1}\sigma^2 \boldsymbol{\Upsilon}_{\boldsymbol{\varphi}}^{-1}\right), \tag{18}$$

holds. We use the notation $\boldsymbol{\Upsilon}_{\boldsymbol{\varphi}} = \mathbb{E}[\boldsymbol{Q}'\boldsymbol{Q}]$ and $\boldsymbol{\Upsilon}_{\boldsymbol{\phi}} = \mathbb{E}[\boldsymbol{Z}'\boldsymbol{Z}]$, where $\boldsymbol{\varphi}$ and $\boldsymbol{\phi}$ signify the relation between the unobserved values $\boldsymbol{u}$,$\boldsymbol{v}$ and $\boldsymbol{y}$ as defined in (8)-(9). These quantities can be estimated consistently by $(1/n)\sum_{i=1}^n \boldsymbol{Q}_i'\boldsymbol{Q}_i$ and $(1/n)\sum_{i=1}^n \boldsymbol{Z}_i'\boldsymbol{Z}_i$, where $\boldsymbol{Q}_i$ [resp. $\boldsymbol{Z}_i$] denotes the ith row of the matrix $\boldsymbol{Q}$ [resp. $\boldsymbol{Z}$]. 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

Since the LAD can be used as an initial estimator for $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$ (Lanne and Saikkonen, 2011) and is found to outperform Student's t-MLE in certain instances (Hecq et al., 2016a), we also present the asymptotic distribution of the model parameters for this estimation method. If ε_t is a sequence of *iid* random variables with mean zero, median zero, finite variance and probability density function $f_\varepsilon(\varepsilon_t;\boldsymbol{\lambda})$ 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. (2016a) and using the model specifications in (15)-(16), it follows that, conditional on the unobserved causal and noncausal components,

$$\sqrt{n}(\boldsymbol{\zeta}_{LAD}-\boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{1}{4f_\varepsilon^2(0)}\boldsymbol{\Upsilon}_{\boldsymbol{\phi}}^{-1}\right), \tag{19}$$

$$\sqrt{n}(\boldsymbol{\xi}_{LAD}-\boldsymbol{\xi}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{1}{4f_\varepsilon^2(0)}\boldsymbol{\Upsilon}_{\boldsymbol{\varphi}}^{-1}\right), \tag{20}$$

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

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 in our simulation study 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):

$$(1 - \phi_1 L)(1 - \varphi_1 L^{-1})y_t - \beta_1 x_{1,t} = \varepsilon_t, \tag{21}$$

where $\phi_1 = 0.3$, $\varphi_1 = 0.5$ and $\beta_1 = 0.3$. The error term ε_t follows a t-distribution with 3 degrees of freedom.[10] $x_{1,t}$ will follow different specifications:

(1) $x_{1,t} \overset{iid}{\sim} t(5, 0)$,

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

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

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

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 $t(5, 0)$ and $\mathcal{N}(0, 1)$ distribution do not generate large 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.

[10] The same simulation study was performed for infinite variance cases, e.g., $\varepsilon_t \overset{iid}{\sim} t(2, 0)$. Similar to results in Hecq et al. (2016a) 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.

		Specification for $x_{1,t}$							
		$x_t \overset{iid}{\sim} t(5,0)$		$x_t \overset{iid}{\sim} \mathcal{N}(0,1)$		$x_t \overset{iid}{\sim} C(0,1)$		$x_t \sim$ AR(1)	
T	Parameter	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
50	ϕ_1	0.313	0.168	0.309	0.168	0.297	0.073	0.291	0.078
	φ_1	0.461	0.164	0.469	0.156	0.491	0.072	0.495	0.065
	β_1	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_1	0.302	0.110	0.301	0.113	0.300	0.039	0.296	0.052
	φ_1	0.489	0.103	0.484	0.106	0.497	0.036	0.499	0.042
	β_1	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_1	0.301	0.037	0.300	0.040	0.300	0.016	0.299	0.022
	φ_1	0.497	0.034	0.498	0.036	0.500	0.008	0.500	0.017
	β_1	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_1	0.300	0.026	0.300	0.025	0.300	0.005	0.299	0.015
	φ_1	0.499	0.023	0.500	0.024	0.500	0.004	0.500	0.013
	β_1	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 1: Finite sample properties of the ML estimator for an MARX(1,1,1) with $\varepsilon_t \overset{iid}{\sim} t(3,0)$

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 theoretical feature, we consider the purely noncausal data generating process as defined in (10) with both ε_t and $x_{1,t} \overset{iid}{\sim} \mathcal{N}(0,1)$. We examine 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.[11] We select the model 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

[11] Due to condition (A5) of Andrews et al. (2006), consistency of Gaussian MLE for the MARX was not shown in Section 4.3 of this paper. However, the consistency of Gaussian MLE for the pure ARX case is established in Hannan et al. (1980). Estimation by OLS yields similar results.

causal, mixed and purely noncausal autoregressive models cannot be identified by Gaussian MLE. Another special case arises when $\varphi_1 = 0$, as there 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_1, 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 et al. (2017) 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.

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.[12] As soon as p is fixed, 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 to what extent both steps are still valid in the MARX framework. To that end, we simulate (21) with $\phi_1 = 0.3$, $\varphi_1 = 0.5$ and $\beta_1 = 0.3$, $\varepsilon_t \overset{iid}{\sim} t(3, 0)$ and $x_{1,t} \overset{iid}{\sim} t(2, 0)$. Purely causal ARX$(p, 1)$ models are estimated by Gaussian MLE, where $p = 0, ..., 4$. 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 exogenous variable $x_{1,t}$, 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 (in %) with which the autoregressive orders p and (r, s) are selected for the MARX model

[12] 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. (2016a).

	$T = 100$			$T = 200$			$T = 500$			$T = 1000$		
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 (in %) with which the autoregressive orders p and (r, s) are 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 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. 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 having estimated pseudo causal models determined by information criteria. We will adapt the lag length (increase p) if necessary.

In a second exercise, we suppose the correct total autoregressive order $p = 2$ is known and investigate the selection of MARX$(r, s, 1)$ and MAR(r, s) models with $r + s = 2$ based on the highest log-likelihood. In Table 5, we observe that the model selection procedure improves with sample size, finding the correct MARX(1,1,1) specification in more than 80% of the cases for $T = 500$ and more than 90% when $T = 1000$. As datasets 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 $T = 100$.

	$T = 100$			$T = 200$			$T = 500$			$T = 1000$		
	(2,0)	(1,1)	(0,2)	(2,0)	(1,1)	(0,2)	(2,0)	(1,1)	(0,2)	(2,0)	(1,1)	(0,2)
MARX	35.5	54.4	10.1	29.0	68.8	2.2	16.9	83.0	0.1	6.8	93.2	0.0
MAR	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 5: Lag-lead order (r, s) selected by highest log-likelihood with fixed $p = 2$ for both MARX and MAR model.

We find that the results of the model selection procedure are sensitive to the values of the parameters chosen in the DGP. Consider for example the case in which $\phi_1 = 0.1$ and $\varphi_1 = 0.7$. This will lead to an overselection of first-order models in the first step of the model selection procedure. Due to the fact that the noncausal parameter value is much larger than the causal one, a noncausal model will be selected in the second step. If we consider $\phi_1 = 0.5$ instead, the overselection of first-order models is less likely in the first step. Hence, it is also more probable that the "correct" mixed causal-noncausal model is eventually chosen in the second step.

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.[13] 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,

[13]IMF Primary Commodity Prices, see `http://www.imf.org/external/np/res/commod/index.aspx`.

- 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 as well as the level of the industrial production index (IP_t). These series are taken from the Federal Reserve Bank of St. Louis database.[14] As commodities are mainly priced in dollar we can expect a contemporaneous negative relation between commodity prices and the U.S. exchange rate. When production increases, more input is needed and this has a positive effect on some of the commodities.

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

5.2 From Expectation Models to MARX

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 but with two regressors (exchange rate and industrial production index):

$$\Delta cp_t = \beta_b \Delta cp_{t-1} + \beta_f \mathbb{E}(\Delta cp_{t+1}|\Omega_t) + \vartheta_1 \Delta s_t + \vartheta_2 \Delta ip_t + u_t. \tag{22}$$

[14] Series name TWEXBMTH and INDPROD at `https://fred.stlouisfed.org`.

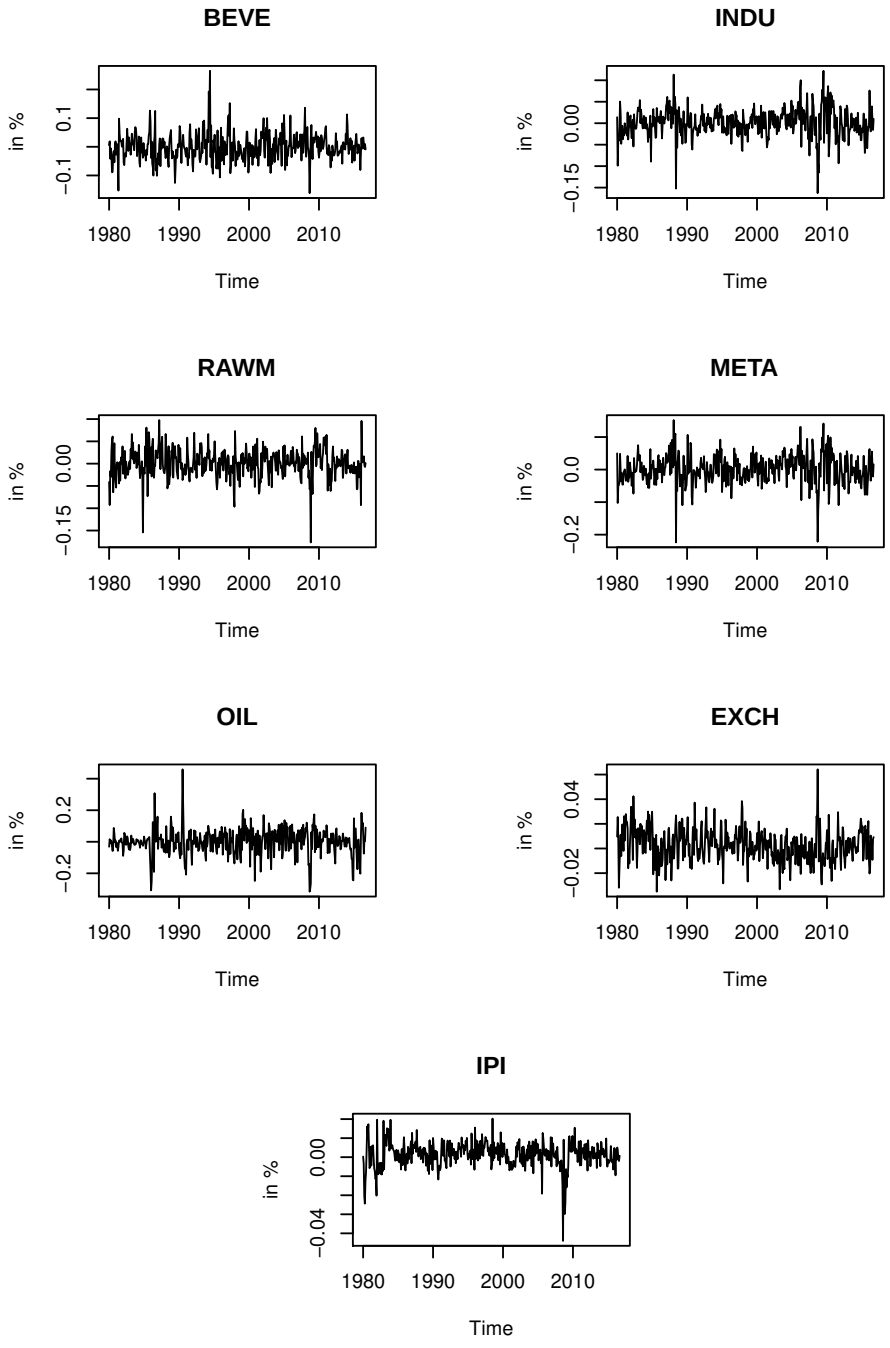

Figure 2: Growth rates of commodity prices, exchange rate and industrial production index

By means of replacing expectations by the future realized growth rate of the commodity price plus an *iid* term ξ_{t+1} and putting the four parts ξ_{t+1}, Δs_t, Δip_t and u_t into the error term, one seemingly obtains an MAR(1,1) with a newly defined disturbance term, say η_t. However, since η_t contains the variable Δip_t and Δs_t, it is likely to be autocorrelated. Lanne and Luoto (2013) assume a MAR$(r-1, s-1)$ structure on η_t to show that the process for Δcp_t can be represented as a MAR(r, s). However, we are not certain that the regressors follow MAR$(r-1, s-1)$ dynamics and besides the direct effect of the fundamentals Δs_t and Δip_t on Δcp_t is lost in this mixed causal-noncausal model. We also do not want to consider a two-step approach as in Lof and Nyberg (2017), hence we prefer rearranging terms as follows

$$\Delta cp_t = \beta_b \Delta cp_{t-1} + \beta_f \Delta cp_{t+1} + \vartheta_1 \Delta s_t + \vartheta_2 \Delta ip_t + \underbrace{u_t + \beta_f \xi_{t+1}}_{\omega_{t+1}}. \tag{23}$$

Now we lag (23) by one period and subsequently divide this equation by β_f to obtain

$$\beta_f^{-1} \Delta cp_{t-1} = \Delta cp_t + \beta_b \beta_f^{-1} \Delta cp_{t-2} + \vartheta_1 \beta_f^{-1} \Delta s_{t-1} + \vartheta_2 \beta_f^{-1} \Delta ip_{t-1} + \beta_f^{-1} \omega_{t+1},$$

which can be rewritten in the following way

$$(1 - \beta_f^{-1} L + \beta_b \beta_f^{-1} L^2) \Delta cp_t = -\vartheta_1 \beta_f^{-1} \Delta s_{t-1} - \vartheta_2 \beta_f^{-1} \Delta ip_{t-1} - \beta_f^{-1} \omega_t. \tag{24}$$

We want to write $(1 - \beta_f^{-1} L + \beta_b \beta_f^{-1} L^2)$ as $(1 - \phi L)(1 - \varphi^* L)$, 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 β_b and β_f, this can be done by taking

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

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 $\varphi = \frac{1}{\varphi^*}$. The polynomial in (24) can be replaced with this result to obtain

$$-\varphi^* L(1-\phi L)(1-\varphi L^{-1})\Delta cp_t = -\vartheta_1\beta_f^{-1}\Delta s_{t-1} - \vartheta_2\beta_f^{-1}\Delta ip_{t-1} - \beta_f^{-1}\omega_t,$$

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

$$(1-\phi L)(1-\varphi L^{-1})\Delta cp_t = \underbrace{\vartheta_1(\varphi^*\beta_f)^{-1}}_{\beta_1}\Delta s_t + \underbrace{\vartheta_2(\varphi^*\beta_f)^{-1}}_{\beta_2}\Delta ip_t + \underbrace{(\varphi^*\beta_f)^{-1}\omega_{t+1}}_{\varepsilon_t}.$$

Since ω_{t+1} is *iid*, we note that ε_t is an *iid* error term. We consequently obtain the MARX(1,1,2) model. Note that we consider a simple one-lag one-lead example from the outset. Introducing more past and future expected terms and more regressors would yield higher order MARX(r,s,q) as solutions. Relaxing the *iid* assumption on ω_{t+1} and allowing it to have an autoregressive MAR$(r-1,s-1)$ structure leads to the same conclusion.

5.3 Identification of Pseudo ARDL models

The first step of our modelling strategy consists in fitting for each $\Delta cp_{i,t}$ series an OLS regression on an intercept, their own lags and the covariates Δs_t and Δip_t:

$$a(L)\Delta cp_{i,t} - b_1\Delta s_t - b_2\Delta ip_t - c = \epsilon_t. \tag{25}$$

The first part of Table 6 reports that step and provides the results concerning models chosen. In all cases, p_{max} is set to 8. The simulation results of Section 4.4 show that BIC tends to underestimate the true order more often than AIC and HQ. Since there is no clear evidence that one of these two performs better than the other, we decide to rely on HQ. However, we always check correlograms and perform LM tests for the null of no autocorrelation and add lags when it is necessary. We put that result, denoted $p(Q)$, in the second row of Table 6. It can be seen that only for the oil commodity series a departure from the result of HQ is necessary to get a white noise. Additionally, we start seaching for autoregressive distributed lag models, i.e., pseudo ARDL$(p_{max}, p_{max}, p_{max})$[15] to verify the inclusion of only the contemporaneous impact of Δs_t and Δip_t. Indeed, as we end up with ARDL$(p_i, 0, 0)$ for all commodity series, our choice is justified. This simple first step can be done using (for instance) EViews ARDL estimation features.[16] Moreover, a simple regression shows that we do not reject the null of linear independence between the two exogenous variables.

The second part of Table 6 reports the estimation results (HCSE robust standard errors in brackets). We observe that commodity prices depend on their own lags as well as on the exchange rate and industrial production (except for BEVE). The highest negative effect of Δs_t is on the OIL index, a result that makes sense given that oil products heavily depend on exports and hence are more negatively influenced by an increase of the USD. The last row of Table 6 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 are able to discriminate components of the MARX model using non-Gaussian MLE. Note that the possibility to identify models by Gaussian MLE or OLS using a strictly exogenous regressor is a feature of purely causal and noncausal models that does not extend to mixed processes. A non-Gaussian MLE is needed for the latter case.

[15] The first argument denotes the amount of lags of the dependent variable, the second and third argument the amount of lags for the exogenous variables.

[16] The same step can be performed in the R package **MARX** by allowing for lags in $\boldsymbol{X}_t$ and comparing information criteria. This approach is however more cumbersome.

	Commodities				
	INDU	META	OIL	BEVE	RAWM
p_{HQ}	2	1	1	1	1
$p(Q)$	2	1	4	1	1
c	0.002	0.002	0.003	0.001	0.001
	(0.0014)	(0.002)	(0.004)	(0.002)	(0.002)
a_1	0.185	0.187	0.298	0.280	0.183
	(0.058)	(0.062)	(0.066)	(0.066)	(0.058)
a_2	0.091		-0.071		
	(0.056)		(0.055)		
a_3			0.027		
			(0.054)		
a_4			-0.122		
			(0.052)		
b_1	-0.889	-1.251	-1.673	-0.587	-0.453
	(0.145)	(0.201)	(0.341)	(0.176)	(0.128)
b_2	0.699	1.048	1.152	0.467	0.604
	(0.229)	(0.288)	(0.558)	(0.275)	(0.313)
$\bar{R}^2$	0.226	0.204	0.181	0.112	0.091
JB	101.33	99.82	132.47	107.64	143.38

Table 6: Estimation results - pseudo causal models

5.4 Identification and Estimation of MARX

Once the number of lags p_i $(i = 1, ..., 5)$ in ARDL models with the contemporaneous Δs_t and Δip_t are determined for each commodity price i, we estimate every possible MARX$(r_i, s_i, 2)$ models with $p_i = r_i + s_i$. We choose the model that gives the highest log-likelihood values. Table 7 reports the final results for each commodity. The values for $\boldsymbol{\phi}$ (resp. $\boldsymbol{\varphi}$) are the estimated coefficients of the lag (resp. lead) polynomials. We can observe some differences in the dynamics of the commodities. It emerges that we have purely causal autoregressive models for commodities INDU, META and RAWM. We have a purely noncausal specification for BEVE and a mixed model for OIL. Distributions are rather leptokurtic: the smallest value for the degrees of freedom parameter is $\hat{\nu} = 2.89$ for OIL, the largest value is $\hat{\nu} = 5.33$ for RAWM. The negative impact

of exchange rate is more pronounced for OIL and the smallest value is for the raw material commodity index. The industrial production index is not significantly different from zero in the OIL and BEVE equations.

	MARX(2,0,2)					
INDU	ϕ_1	0.247 (0.038)	β_1	-0.722 (0.089)	c	0.001 (0.002)
	ϕ_2	0.096 (0.038)	β_2	0.764 (0.175)	$[\nu, \sigma]$	[4.165, 0.021]
	MARX(1,0,2)					
META	ϕ_1	0.265 (0.038)	β_1	-1.062 (0.131)	c	0.002 (0.002)
			β_2	0.992 (0.255)	$[\nu, \sigma]$	[5.299, 0.032]
	MARX(1,0,2)					
RAWM	ϕ_1	0.191 (0.041)	β_1	-0.374 (0.098)	c	0.001 (0.001)
			β_2	0.457 (0.191)	$[\nu, \sigma]$	[5.338, 0.024]
	MARX(2,2,2)					
OIL	ϕ_1	-0.491 (0.036)	β_1	-1.100 (0.212)	c	0.007 (0.003)
	ϕ_2	-0.176 (0.036)	β_2	0.284 (0.413)	$[\nu, \sigma]$	[2.890, 0.048]
	φ_1	0.725 (0.034)				
	φ_2	-0.241 (0.034)				
	MARX(0,1,2)					
BEVE	φ_1	0.295 (0.039)	β_1	-0.409 (0.141)	c	-0.001 (0.001)
			β_2	0.241 (0.276)	$[\nu, \sigma]$	[4.954, 0.034]

Table 7: Estimation results - MARX models

Since 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, the necessity of exogenous variables in a noncausal model could be questioned. Significance of the exchange rate in every series and the production index in three commodities in this empirical application seems to indicate that not all of its effect is predictable by the respective commodity price dynamics. Indeed, estimating MAR(r, s), namely when excluding the exogenous regressors, we obtain MAR(1,1) for INDU, MAR(0,1) for BEVE, MAR(1,0) for META, MAR(0,4) for OIL and MAR(0,1) for RAWM. This leads to having a noncausal component in each series but META and illustrates that noncausal models can indeed capture the information that economic agents have but not the econometrician. Explictly adding regressors gives different models and hence

justifies the use of the MARX(r, s, q) models both for forecasting and for the understanding of economic relationships. Lastly, we can also illustrate the identification feature raised in Section 2.2. Let us consider the META commodity as an example, as we obtained a purely causal model and coefficients for both exogenous regressors which are high and significantly different from zero. If we now estimate a purely causal MARX(1,0,2) and purely noncausal MARX(0,1,2) by OLS, we obtain values for the log-likelihood of respectively 788.65 and 785.59, which indicate that the causal pattern is favored.

6 Conclusion

This paper proposes to estimate mixed causal-noncausal models by non-Gaussian MLE when additional regressors are present. We have in mind the estimation of structural relationships subject to expectation schemes such as the new Hybrid Phillips curve or lag-augmented present value models. Many empirical macroeconomic equations are covered by this framework. We provide a successful empirical illustration on the relation between commodity prices, the exchange rate and the industrial production index.

The one-step approach to estimating MARX is easy to implement and the estimation of the standard errors that we propose is quite robust to computational overflows. It allows to estimate directly the impact of exogenous variables without the need to augment the MAR with leads and lags (and to lose the impact of $\boldsymbol{X}_t$) or to use a two-step approach as in Lof and Nyberg (2017). In addition, we find that the presence of strictly exogenous regressors have the appealing feature that one can discriminate between purely causal and noncausal specifications in a Gaussian framework.

References

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

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

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.

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

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

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

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

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

Cubadda, G., Hecq, A. and S. Telg (2017), Serial Correlation Common Noncausal Features, *MPRA Paper 77254*, University Library of Munich, Germany.

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

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

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

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.

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

Hecq, A., Lieb, L. and S. Telg (2016a), Identification of Mixed Causal-Noncausal Models in Finite Samples, *Annals of Economics and Statistics*, 123/124, 307-331.

Hecq, A. Lieb, L. and S. Telg (2017), 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. and L. Lieb (2016b), Do Seasonal Adjustments Induce Noncausal Dynamics in Inflation Rates?, *MPRA Paper 74922*, University Library of Munich, Germany.

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.

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

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

Lanne, M. and J. Luoto (2013), Autoregression-Based Estimation of the New Keynesian Phillips Curve, *Journal of Economic Dynamics & Control*, 37, 561-570.

Lanne, M., Nyberg, H. and E. Saarinen (2012b). Does Noncausality Help in Forecasting Economic Time Series?, *Economics Bulletin*, 32(4), 2849-2859.

Lanne, M. and P. Saikkonen (2011), Noncausal Autoregressions for Economic Time Series, *Journal of Time Series Econometrics*, 3(3), 1-32.

Lanne, M. and P. Saikkonen (2013), Noncausal Vector Autoregression, *Econometric Theory*, 29(3), 447-481.

Lof, M. and H. Nyberg (2017), Noncausality and the Commodity Currency Hypothesis, *Energy Economics*, 65, 424-433.

Pesaran, H. (2015), *Time Series and Panel Data Econometrics*, Oxford University Press.

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 consider a single explanatory variable denoted x_t^*. The transfer function model is given by

$$y_t = \psi^*(L)x_t^* + n_t, \tag{26}$$

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 (26) by assuming that the transfer function operator

can be expressed in a rational factorization as $\psi^*(z) = a(z)^{-1}\theta^*(z)$. Multiplying (26) by $a(L)$ yields

$$\begin{aligned} a(L)y_t &= \theta^*(L)x_t^* + a(L)n_t \\ &= \theta^*(L)x_t^* + \varepsilon_t^*, \end{aligned} \tag{27}$$

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

$$\begin{aligned} \phi(L)\varphi(L^{-1})y_t &= \left(-\frac{1}{\varphi_s^*} + \frac{\theta_1^*}{\varphi_s^*} + ...\frac{\theta_k^*}{\varphi_s^*}\right) x_{t+s}^* - \frac{1}{\varphi_s^*}\varepsilon_{t+s}^* \\ &= \theta(L)x_t + \varepsilon_t. \end{aligned} \tag{28}$$

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

Part B - Approximate Likelihood Function

Define $\boldsymbol{b} = \boldsymbol{\beta}'\tilde{\boldsymbol{x}}$ such that $\boldsymbol{z} = \boldsymbol{BAy} - \boldsymbol{\beta}'\tilde{\boldsymbol{x}} = \boldsymbol{BAy} - \boldsymbol{b}$ (See case 5 in the proof of Lemma 1 in Appendix C). Assume $\boldsymbol{B}$ and $\boldsymbol{A}$ are invertible. We are interested in the inverse transformation, i.e. $\boldsymbol{y} = \boldsymbol{Q}(\boldsymbol{z} + \boldsymbol{b})$, where $\boldsymbol{Q} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$. Let $\boldsymbol{Q}$ be a (2×2) matrix, then we have

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} q_1 & q_2 \\ q_3 & q_4 \end{bmatrix} \left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \right), \tag{29}$$

with the following functions

$$y_1 = g_1(z_1, z_2) = q_1 z_1 + q_2 z_2 + b_1, \tag{30}$$

$$y_2 = g_2(z_1, z_2) = q_3 z_1 + q_4 z_2 + b_2. \tag{31}$$

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

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} \\ \frac{\partial y_2}{\partial z_1} & \frac{\partial y_2}{\partial z_2} \end{bmatrix} = \begin{bmatrix} q_1 & q_2 \\ q_3 & q_4 \end{bmatrix} = \boldsymbol{Q}. \tag{32}$$

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

$$f_{y_1,y_2}(y_1, y_2) = \frac{1}{|\det(\boldsymbol{Q})|} f_{z_1,z_2}(\boldsymbol{BAy} - \boldsymbol{b}) \tag{33}$$

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 (35)-(36) and $\boldsymbol{Q} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$, we find that the probability density of the process y_t can be represented in the following way:

$$f_{y;\boldsymbol{\lambda}}(\boldsymbol{y}) = \frac{1}{|\det(\boldsymbol{Q})|} f_z(\boldsymbol{BAy} - \boldsymbol{b}; \boldsymbol{\lambda})$$

$$
\begin{aligned}
&= |\det(\boldsymbol{A})||\det(\boldsymbol{B})| h_V(\boldsymbol{BAy}-\boldsymbol{b}) f_\varepsilon(\boldsymbol{BAy}-\boldsymbol{b};\boldsymbol{\lambda}) h_U(\boldsymbol{BAy}-\boldsymbol{b}) \\
&= |\det(\boldsymbol{A})| h_V(\boldsymbol{BAy}-\boldsymbol{b}) \left(\prod_{t=r+1}^{T-s} f_\sigma(\boldsymbol{BAy}-\boldsymbol{b};\boldsymbol{\lambda}) \right) h_U(\boldsymbol{BAy}-\boldsymbol{b}) \\
&= h_V(\varphi(L^{-1})y_1, ..., \varphi(L^{-1})y_r) \left(\prod_{t=r+1}^{T-s} f_\sigma(\phi(L)\varphi(L^{-1})y_t - \boldsymbol{\beta}'\boldsymbol{X}_t;\boldsymbol{\lambda}) \right) \\
&\quad h_U(\phi(L)y_{T-s+1}, ..., \phi(L)y_T)|\det(\boldsymbol{A})|, \qquad (34)
\end{aligned}
$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are two nonsingular matrices with $\det(\boldsymbol{B}) = 1$; 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 $\det(\boldsymbol{A})$ is independent of sample size, the density of y_t can be approximated by the second term in (34).

6.1 Part C - Proofs

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

Note that by the definition of u in equation (8), (ii') $(y_1, ..., y_r, u_{r+1}, ..., u_T) =$ $(y_1, ..., y_r, \phi(L)y_{r+1}, ..., \phi(L)y_T)$. 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 by the definition of v in equation (9), (iii') $(v_1, ..., v_{T-s}, y_{T-s+1}, ..., y_T) =$

$(\varphi(L^{-1})y_1, ..., \varphi(L^{-1})y_{T-s}, y_{T-s+1}, ..., y_T)$. 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 (as all information sets are augmented with the same information).

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

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

$(y_1, ..., y_r, \varphi(L^{-1})u_{r+1} - \boldsymbol{\beta}'\boldsymbol{X}_{r+1}, ..., \varphi(L^{-1})u_{T-s} - \boldsymbol{\beta}'\boldsymbol{X}_{T-s}, u_{T-s+1}, ..., u_T)$ by using the second equality in equation (8). Since $\varepsilon_{T-s} = u_{T-s} - \varphi_1 u_{T-s+1} - ... - \varphi_s u_T - \boldsymbol{\beta}'\boldsymbol{X}_{T-s}$ with $u_{T-s+1}, ..., u_T$, $\boldsymbol{X}_{T-s}$ 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) $(v_1, ..., v_r, \varepsilon_{r+1}, ...\varepsilon_{T-s}, y_{T-s+1}, .., y_T) =$

$(v_1, ..., v_r, \phi(L)v_{r+1} - \boldsymbol{\beta}'\boldsymbol{X}_{r+1}, ..., \phi(L)v_{T-s} - \boldsymbol{\beta}'\boldsymbol{X}_{T-s}, y_{T-s+1}, ..., y_T)$ by using the second equality in equation (9). Since $\varepsilon_{r+1} = v_{r+1} - \phi_1 v_r - ... - \phi_r v_1 - \boldsymbol{\beta}'\boldsymbol{X}_{r+1}$ with $v_1, ..., v_r$, $\boldsymbol{X}_{r+1}$ 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: $(y_1, ..., y_T, \boldsymbol{X}_r, ..., \boldsymbol{X}_{T-s}) \sim (v_1, ..., v_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, u_{T-s+1}, ..., u_T)$. Denote the vector corresponding to the first information set by $\tilde{\boldsymbol{y}}$ and the second one by $\boldsymbol{z}$. This statement can be proven using the algebra in Lanne and Saikkonen (2011). Define the vectors $\boldsymbol{w} =$

$[v_1, ..., v_{T-s}, u_{T-s+1}, ..., u_T]'$ and $\boldsymbol{y} = [y_1, ..., y_T]'$. Then

$$\begin{bmatrix} v_1 \\ \vdots \\ v_{T-s} \\ u_{T-s+1} \\ \vdots \\ v_T \end{bmatrix} = \begin{bmatrix} y_1 - \varphi_1 y_2 - ... - \varphi_s y_{s+1} \\ \vdots \\ y_{T-s} - \varphi_1 y_{T-s+1} - ... - \varphi_s y_T \\ y_{T-s+1} - \phi_1 y_{T-s} - ... - \phi_r y_{T-s+1-r} \\ \vdots \\ y_T - \phi_1 y_{T-1} - ... - \phi_r y_{T-r} \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} y_1 \\ \vdots \\ y_{T-s} \\ y_{T-s+1} \\ \vdots \\ y_T \end{bmatrix}, \tag{35}$$

which can be written as $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{y}$. Now define $\tilde{\boldsymbol{x}} = [\underbrace{0, ..., 0}_{r \text{ times}}, \boldsymbol{X}_{r+1}, ..., \boldsymbol{X}_{T-s}, \underbrace{0, ..., 0}_{s \text{ times}}]'$. Similarly, we can form the following system of equations (with slight abuse of notation, as $\boldsymbol{X}_t$ is a vector) for the vector $\boldsymbol{z}$. That is, equation (36) shows that $\boldsymbol{z} = \boldsymbol{B}\boldsymbol{w} - \boldsymbol{\beta}'\tilde{\boldsymbol{x}}$. Combining both systems of equations, we find that the vectors $\boldsymbol{z}$ and $\tilde{\boldsymbol{y}}$ are related in the following way: $\boldsymbol{z} = \boldsymbol{B}\boldsymbol{A}\boldsymbol{y} - \boldsymbol{\beta}'\tilde{\boldsymbol{x}}$, where $\boldsymbol{y}$ and $\tilde{\boldsymbol{x}}$ combined form the information set $\tilde{\boldsymbol{y}}$. Since the matrices $\boldsymbol{B}$ and $\boldsymbol{A}$ as well as the parameter vector $\boldsymbol{\beta}$ only contain the known parameters, this shows that these information sets are equivalent. Combining all cases shows that information sets $(i) - (vi)$ are equivalent.

$$\begin{bmatrix} v_1 \\ \vdots \\ v_r \\ v_{r+1} - \phi_1 v_r - ... - \phi_r v_1 - \boldsymbol{\beta}'\boldsymbol{X}_{r+1} \\ \vdots \\ v_{T-s} - \phi_1 v_{T-s-1} - ... - \phi_r v_{T-s-r} - \boldsymbol{\beta}'\boldsymbol{X}_{T-s} \\ u_{T-s+1} \\ \vdots \\ u_T \end{bmatrix} = \boldsymbol{B} \begin{bmatrix} v_1 \\ \vdots \\ v_r \\ v_{r+1} \\ \vdots \\ v_{T-s} \\ u_{T-s+1} \\ \vdots \\ u_T \end{bmatrix} - \boldsymbol{\beta}' \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \boldsymbol{X}_{r+1} \\ \vdots \\ \boldsymbol{X}_{T-s} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{36}$$

Proof of Lemma 2

For the proof of this lemma and the next theorem, we need some additional notation. Define $e_t = \frac{f'_\sigma(\varepsilon_t;\boldsymbol{\lambda})}{f_\sigma(\varepsilon_t;\boldsymbol{\lambda})} = \frac{f'(\varepsilon_t/\sigma;\boldsymbol{\lambda})}{\sigma f(\varepsilon_t/\sigma;\boldsymbol{\lambda})}$, $\tilde{\mathcal{J}} = \sigma^{-2}\mathcal{J}$, $\tilde{\mathcal{I}} = \sigma^{-2}\mathcal{I}$ and $n \equiv (T-p)$. Furthermore, let $x = \varepsilon_t/\sigma$, then we have that

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

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

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

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

$$\mathbb{E}(z_s e_t) = \begin{cases} 0, & \text{if } s \neq t, \\ -1, & \text{if } s = t, \end{cases} \tag{37}$$

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

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

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

Hence, we note that $\boldsymbol{V}_{t-1}$ and e_t are still independent as in Lanne and Saikkonen (2011a). Consequently, we still find

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

where $\boldsymbol{\Gamma}_V$ denotes the autocovariance matrix of the vector $\boldsymbol{V}_{t-1}$. Because $\boldsymbol{V}_{t-1}e_t$ is uncorrelated, we have

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

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

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

That is, the independence of e_t and $\boldsymbol{U}_{t+1}$ is preserved through strict exogeneity. Letting $\boldsymbol{\Gamma}_U$ be the autocovariance matrix of $\boldsymbol{U}_{t+1}$, we find that

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

$$= \mathbb{E}(e_t^2)\mathbb{E}(\boldsymbol{U}_{t+1}\boldsymbol{U}'_{t+1})$$
$$= \tilde{\mathcal{J}}\boldsymbol{\Gamma}_U.$$

Because $\boldsymbol{U}_{t+1}e_t$ is uncorrelated, we have

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

Lastly, we can apply the same logic for the parameter vector $\boldsymbol{\beta}$. Since for $i = 1, ..., q$, we have that

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

by the independence of $x_{i,t}$ and ε_t. If we denote by $\boldsymbol{\Gamma}_X$, the autocovariance matrix of $\boldsymbol{X}_t$, it follows that

$$\text{Cov}\left(\frac{\partial g_t(\boldsymbol{\theta_0})}{\partial \boldsymbol{\beta}}\right) = \text{Cov}(-\boldsymbol{X}_t e_t)$$
$$= \mathbb{E}(e_t^2)\mathbb{E}(\boldsymbol{X}_t\boldsymbol{X}'_t)$$
$$= \tilde{\mathcal{J}}\boldsymbol{\Gamma}_X.$$

Because $\boldsymbol{X}_t e_t$ is uncorrelated, we have

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

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

$$\text{Cov}(z_{t-i}e_t, z_{k-j}e_k) = \begin{cases} \mathcal{I} + \tilde{\mathcal{J}} \sum_{m=1}^{q} \beta_m^2 \sigma_{x_m}^2 & \text{if } t = k, i = j = 0, \\ \mathcal{J} + \tilde{\mathcal{J}} \sum_{m=1}^{q} \beta_m^2 \sigma_{x_m}^2 & \text{if } t = k, i = j \neq 0, \\ 1 & \text{if } t \neq k, i = t-k, j = k-t, \\ 0 & \text{otherwise.} \end{cases} \tag{38}$$

Hence, using (37)-(38), 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 $\partial g_t(\boldsymbol{\theta_0})/\partial \boldsymbol{\phi}$ and $\partial g_t(\boldsymbol{\theta_0})/\partial \boldsymbol{\varphi}$, first consider for $1 \leq i \leq r, 1 \leq j \leq s$:

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

$$= \begin{cases} \alpha_{t-k-i}\delta_{t-k-j}, & \text{for } t > k, \\ \tilde{\mathcal{J}} \sum_{a=0}^{\infty} \alpha_a \sum_{m=1}^{q} \beta_m^2 Q_m(i,j,a) & \text{for } t = k, \\ 0 & \text{for } t < k. \end{cases}$$

The element (i,j) of the matrix $\frac{1}{n}\text{Cov}(\partial L_T(\boldsymbol{\theta_0})/\partial\boldsymbol{\phi}, \partial L_T(\boldsymbol{\theta_0})/\partial\boldsymbol{\varphi})$ is

$$\begin{aligned}
& n\text{Cov}\left(\frac{1}{n}\sum_{t=r+1}^{T-s}\frac{\partial g_t(\boldsymbol{\theta_0})}{\partial\phi_i}, \frac{1}{n}\sum_{k=r+1}^{T-s}\frac{\partial g_k(\boldsymbol{\theta_0})}{\partial\varphi_j}\right) \\
&= \frac{1}{n}\sum_{t=r+1}^{T-s}\sum_{k=r+1}^{T-s}\text{Cov}\left(\frac{\partial g_t(\boldsymbol{\theta_0})}{\partial\phi_i}, \frac{\partial g_k(\boldsymbol{\theta_0})}{\partial\varphi_j}\right) \\
&= \frac{1}{n}\sum_{t=r+1}^{T-s}\sum_{k=r+1}^{T-s}\text{Cov}(-v_{t-i}e_t, -u_{k+j}e_k) \\
&= \frac{1}{n}\sum_{t=r+1}^{T-s}\sum_{k=r+1}^{T-s}\left(\mathbb{1}_{\{t>k\}}\alpha_{t-k-i}\delta_{t-k-j} + \mathbb{1}_{\{t=k\}}\tilde{\mathcal{J}}\sum_{a=0}^{\infty}\alpha_a\sum_{m=1}^{q}\beta_m^2 Q_m(i,j,a)\right) \\
&= \frac{1}{n}\left(\sum_{k=r+1}^{T-s-1}\sum_{t=k+1}^{T-s}\alpha_{t-k-i}\delta_{t-k-j} + n\tilde{\mathcal{J}}\sum_{a=0}^{\infty}\alpha_a\sum_{m=1}^{q}\beta_m^2 Q_m(i,j,a)\right) \\
&= \frac{1}{n}\left(\sum_{k=r+1}^{T-s-1}\sum_{t=0}^{T-s-k-i}\alpha_t\delta_{t+i-j}\right) + \tilde{\mathcal{J}}\sum_{a=0}^{\infty}\alpha_a\sum_{m=1}^{q}\beta_m^2 Q_m(i,j,a) \\
&\to \sum_{t=0}^{\infty}\alpha_t\delta_{t+i-j} + \tilde{\mathcal{J}}\sum_{a=0}^{\infty}\alpha_a\sum_{m=1}^{q}\beta_m^2 Q_m(i,j,a),
\end{aligned}$$

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 $\boldsymbol{\phi}$ and the parameter vector of the exogenous

variables $\boldsymbol{\beta}$. That is,

$$\begin{aligned}\operatorname{Cov}\left(\frac{\partial g_t(\boldsymbol{\theta}_0)}{\partial \phi_i}, \frac{\partial g_k(\boldsymbol{\theta}_0)}{\partial \beta_j}\right) &= \operatorname{Cov}(-v_{t-i}e_t, -x_{j,k}e_k) \\ &= \mathbb{E}\left(\sum_{a=0}^{\infty} \alpha_a \varepsilon_{t-i-a} x_{j,k} e_k e_t\right) \\ &+ \mathbb{E}\left(\sum_{a=0}^{\infty} \alpha_a \sum_{m=1}^{q} \beta_m x_{m,t-i-a} x_{j,k} e_k e_t\right) \\ &= \begin{cases} \beta_j \tilde{\mathcal{J}} \sum_{a=0}^{\infty} \alpha_a \gamma_{x_j}(i+a) & \text{for } t = k, 1 \leq i \leq r, \\ 0 & \text{for } t \neq k, 1 \leq i \leq r. \end{cases}\end{aligned}$$

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 $\boldsymbol{\varphi}$ and the parameter vector of the exogenous variables $\boldsymbol{\beta}$:

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

The proof of asymptotic normality is similar to Breidt et al. (1991) and Lanne and Saikkonen (2011). Define $\boldsymbol{M} = \operatorname{diag}(\boldsymbol{\Sigma}, \boldsymbol{\Omega})$, $\boldsymbol{W}_t = \frac{\partial g_t(\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta}}$ and note that $n \equiv (T-p)$. By the Cramér-Wold theorem, it suffices to show that for any vector $\boldsymbol{a}$ of appropriate size,

$$\frac{1}{\sqrt{n}} \sum_{t=r+1}^{T-s} \boldsymbol{a}' \boldsymbol{W}_t \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{a}' \boldsymbol{M} \boldsymbol{a}). \tag{39}$$

Define the sequence of $(p+q+d+1)$ dimensional random vectors $\{\boldsymbol{W}_{tm}, t \in \mathbb{Z}\}$ to be the partials defined in Section 3.1, where v_t, u_t and all $x_{i,t}$ for $i = 1, ..., q$ are replaced by their representation in (11)-(12) and assumption (A2) with the sums truncated at a large positive integer m, i.e.,

$$v_t^{(m)} = \sum_{j=0}^{m} \alpha_j z_{t-j},\ u_t^{(m)} = \sum_{j=0}^{m} \delta_j z_{t+j} \text{ and } x_{i,t}^{(m)} = c_i + \sum_{j=-m}^{m} \rho_{i,j} \eta_{i,t-j}.$$

It can be verified that $\mathbb{E}(\boldsymbol{W}_t) = \boldsymbol{0}$ and $\boldsymbol{\gamma}_{\boldsymbol{W}_t}(0) + 2\sum_{j=1}^{\infty} \boldsymbol{\gamma}_{\boldsymbol{W}_t}(j) \neq \boldsymbol{0}$. This result also holds for $\boldsymbol{W}_t$ replaced by $\boldsymbol{W}_{tm}$. Let $\boldsymbol{M}_m$ be the matrix corresponding to $\boldsymbol{M}$, obtained by truncating u_t, v_t and $\boldsymbol{X}_t$. Then the stationary sequence $\{\boldsymbol{W}_{tm}, t \in \mathbb{Z}\}$ is $\max\{m+p, 2m\}$ dependent.[17] Now that we verified the conditions, we can apply Theorem 6.4.2 in Brockwell and Davis (1991) to obtain

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

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

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

the convergence in (39) is immediate from Proposition 6.3.9 in Brockwell and Davis (1991). The positive definiteness of $\boldsymbol{\Sigma}$ can be established similar to the proof in Breidt et al. (1991). In the MARX case, the block matrix $\boldsymbol{\Sigma}$ is given as

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

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

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

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

where $\boldsymbol{A}$ is $r \times r$, $\boldsymbol{C}$ is $s \times s$ and $\boldsymbol{B}$ is $r \times s$. Consider the vectors $\boldsymbol{P} = [P_1, ..., P_r]'$ and $\boldsymbol{S} = [S_1, ..., S_s]'$ defined by

$$\boldsymbol{P}_t = \sum_{a=0}^{\infty} \alpha_a z_{-t-a} e_0, \qquad \text{for } t = 1, ..., r, \tag{40}$$

$$\boldsymbol{S}_t = \sum_{b=0}^{\infty} \delta_b z_{t-b} e_0, \qquad \text{for } t = 1, ..., s. \tag{41}$$

It can easily be verified that the covariance matrices of $\boldsymbol{P}$ and $\boldsymbol{S}$, denoted $\boldsymbol{\Sigma}_{PP}$ and $\boldsymbol{\Sigma}_{SS}$, are equal to $\boldsymbol{A}$ and $\boldsymbol{C}$. From (38), 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 also 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 $\boldsymbol{A}$ and $\boldsymbol{C}$ are positive definite since there is no linear dependence within the vectors $\boldsymbol{P}$ and $\boldsymbol{S}$. We proceed by proving the positive definiteness of $\tilde{\boldsymbol{\Sigma}}$ by showing that the Schur Complement of the block $\boldsymbol{A}$ of the matrix $\tilde{\boldsymbol{\Sigma}}$ given as $\boldsymbol{C} - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B}$ is positive definite. We know that the covariance matrix of $\boldsymbol{S} - \boldsymbol{\Sigma_{SP}}\boldsymbol{\Sigma_{PP}}^{-1}$, i.e. $\boldsymbol{C} - (\mathcal{J} + \tilde{\mathcal{J}} \sum_{m=1}^{q} \beta_m^2 \sigma_m^2)\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B}$, is positive semidefinite and hence for a nonzero vector

$\boldsymbol{c} \in \mathbb{R}^s$ with $\boldsymbol{Bc} \neq \boldsymbol{0}$, we have that

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

Alternatively, if $\boldsymbol{Bc} = \boldsymbol{0}$, we have that

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

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

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

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

Proof of Theorem 1

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

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

which can easily be verified using the quotient rule. Let $\boldsymbol{Y}_t$ 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 $\tilde{\boldsymbol{V}}_{t-1} = [\tilde{v}_{t-1}, ..., \tilde{v}_{t-r}]'$ and

$\tilde{\boldsymbol{U}}_{t+1} = [\tilde{u}_{t+1}, ..., \tilde{u}_{t+s}]'$ 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 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{\boldsymbol{V}}_{t-1}\tilde{\boldsymbol{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{\boldsymbol{U}}_{t+1}\tilde{\boldsymbol{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})\boldsymbol{X}_t\boldsymbol{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{\boldsymbol{V}}_{t-1}\tilde{\boldsymbol{U}}_{t+1}' + \sigma^{-1}h(\sigma^{-1}\tilde{\varepsilon}_t;\boldsymbol{\lambda})\boldsymbol{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{\boldsymbol{V}}_{t-1}\boldsymbol{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{\boldsymbol{V}}_{t-1} + \sigma^{-2}h(\sigma^{-1}\tilde{\varepsilon}_t;\boldsymbol{\lambda})\tilde{\boldsymbol{V}}_{t-1}, \\
\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\phi}\partial\boldsymbol{\lambda}' &= -\sigma^{-1}\tilde{\boldsymbol{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{\boldsymbol{U}}_{t+1}\boldsymbol{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{\boldsymbol{U}}_{t+1} + \sigma^{-2}h(\sigma^{-1}\tilde{\varepsilon}_t;\boldsymbol{\lambda})\tilde{\boldsymbol{U}}_{t+1}, \\
\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\varphi}\partial\boldsymbol{\lambda}' &= -\sigma^{-1}\tilde{\boldsymbol{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\boldsymbol{X}_t + \sigma^{-2}h(\sigma^{-1}\tilde{\varepsilon}_t;\boldsymbol{\lambda})\boldsymbol{X}_t, \\
\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\lambda}' &= -\sigma^{-1}\boldsymbol{X}_t\partial h(\sigma^{-1}\tilde{\varepsilon}_t;\boldsymbol{\lambda})/\partial\boldsymbol{\lambda}', \\
\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}'.
\end{aligned}
$$

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

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

where $\boldsymbol{c}\in\mathbb{R}^{p+q+1+d}$ and the argument $\boldsymbol{\theta}^*_{\boldsymbol{T}}(\boldsymbol{c})$ 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 $T^{-1/2}\boldsymbol{c}$. If $\|\cdot\|$ denotes the Euclidian norm we have $\sup_{\boldsymbol{c}\in\boldsymbol{C}}\left\|\boldsymbol{\theta}^*_{\boldsymbol{T}}(\boldsymbol{c})-\boldsymbol{\theta_0}\right\|\to 0$ for any compact set $\boldsymbol{C}\subset\mathbb{R}^{p+q+1+d}$. Using the dominance conditions (A7) in Davis et al. (1992), arguments similar to Breidt et al. (1991, p. 186-190) and assumption (A1) in this paper, it can be shown that a uniform law of large numbers for stationary ergodic processes applies to $\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ over any small enough compact neighborhood $\boldsymbol{\theta_0}$. We can conclude that

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

for $\boldsymbol{c}\in\boldsymbol{C}$. 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.