Observation-Driven filters for Time-Series with Stochastic Trends and Mixed Causal Non-Causal Dynamics*

Francisco Blasques[(a,b)], Siem Jan Koopman[(a,b)], Gabriele Mingoli[(a,b,*)]

(a) Vrije Universiteit Amsterdam, (b) Tinbergen Institute.

March 7, 2024

PRELIMINARY DRAFT. PLEASE DO NOT QUOTE.

Abstract

This paper proposes a novel time-series model with a non-stationary stochastic trend, locally explosive mixed causal non-causal dynamics and fat-tailed innovations. The model allows for a description of financial time-series that is consistent with financial theory, for a decomposition of the time-series in trend and bubble components, and for meaningful real-time forecasts during bubble episodes. We provide sufficient conditions for strong consistency and asymptotic normality of the maximum likelihood estimator. The model-based filter for extracting the trend and bubbles is shown to be invertible and the extracted components converge to the true trend and bubble paths. A Monte Carlo simulation study confirms the good finite sample properties. Finally, we consider an empirical study of Nickel monthly price series and global mean sea level data. We document the forecasting accuracy against competitive alternative methods and conclude that our model-based forecasts outperform all these alternatives.

Key words: observation-driven filter, non-stationary time-series, mixed causal non-causal models.

*Corresponding author: Gabriele Mingoli. Email: g.mingoli@vu.nl.

F. Blasques is thankful to the Dutch Science Foundation (NWO) for financial support (VI.Vidi.195.099).

1 Introduction

Financial and economic time series often experience periods of locally explosive behaviour that are followed by strong and sharp decline or mean-reverting dynamics. These events, often called bubbles, are the object of much attention and discussion, especially by those engaged in trading financial assets and commodities. Bubbles in housing markets and stock markets have arguably been either the cause or an early symptom of emerging widespread global economic recessions.

The economic literature as often described bubbles in asset prices as sum between a *fundamental value* which is founded on rational expectations and a locally explosive component, which is referred to as a *speculative bubble*, see e.g. Blanchard and Watson (1982)and West (1987). Empirically, Diba and Grossman (1988) proposed the use of unit root and cointegration tests to test for explosive non-stationary bubbles in the data. This testing methodology relies however on the time-series being globally non-stationary and/or explosive in nature. This assumption can be problematic since locally explosive processes featuring short-lived bubbles which expand and collapse may render the time-series stationary Evans (1991). This concern is addressed by Phillips et al. (2011), Phillips et al. (2015), Phillips and Shi (2018) which develop supremum tests on recursive right-side unit root test-statistics in order to allow for exploding sub-samples in a time-series. Interestingly, these methods also allow for the dating of both the beginning and the end of explosive bubble events. Further, Homm and Breitung (2012) finds that this recursive method works well as a real-time bubble detection algorithm. In empirical applications, (Phillips et al., 2011; Phillips and Yu, 2011) find evidence of bubble events on the Nasdaq index, the U.S. housing price index, the price of crude oil and the spread between Baa and Aaa bond rates. The generealized test proposed in Phillips et al. (2015) was used to detect bubbles in commodity prices (Etienne et al., 2014; Gutierrez, 2013) and real estate (Chen and Funke, 2013; Yiu and Jin, 2012), among others. Recent empirical applications of these same testing methodologies can be found on balooning sovereign risk (Phillips and Shi, 2019), sector trading in real time (Milunovich et al., 2019) and the U.S. regional housing market (Shi, 2017).

In parallel, we have witnessed the emergence of the literature on mixed causal non-causal autoregressive (MAR) models as an alternative form of describing and modeling bubbles in financial and economic time-series. As such MAR models are relevant in the study of financial bubbles. These models have gained considerable attention in the last decade as they have proven able to fit a number of interesting episodes in financial data and allowed for some modeling structure and formalization of the concept of a locally explosive event, or bubble, at least in a time-series sense.

Mixed causal non-causal models allow for speculative bubble dynamics to be modeled using a noncausal autoregressive process of order one with heavy tailed innovations. This specification is able to model speculative bubbles since it generates large outliers which are preceded by a slow build-up to that outlying observation. The class

of MAR models was recently extended to accommodate for stable distributions by Gouriéroux and Zakoïan (2017) and a higher order mixed causal and noncausal polynomial structure by Fries and Zakoïan (2017). The MAR framework has been used to model and forecast financial bubbles in a wide range of different assets prices displaying explosive behaviors like Nickel monthly price, NASDAQ price, Bitcoin price, and the price of a number of different commodities (Hecq and Voisin (2021), Fries and Zakoïan (2017), Hencic and Gouriéroux (2015), Gouriéroux and Zakoïan (2017)).

Despite the numerous applications of MAR models, it is crucial to note that these models operate in a stationary framework. The underlying stationarity of MAR processes allows for the derivation of theoretical results, such as expected bubble life times and emergence and collapse probabilities. Of course, this also means that unit root tests applied to MAR data generating processes will generally reject the unit root hypothesis and that MAR models are unable to distinguish the potential speculative bubble from the fundamental value of a financial asset. In practice, a fundamental challenge faced by MAR applications is indeed that most financial time-series displaying locally explosive behaviours are also non-stationary. The fact that most financial time-series feature some form of non-stationary fundamental component, typically exhibiting unit-root or random-walk dynamics, is practically undisputed. Rather, it is the challenge in appropriately dealing with the non-stationary component which has made it difficult for MAR models to be applied to non-stationary data.

In the literature for MAR models the recognition that most relevant time-series are non-stationary has led to the use of several detrending methods, the most common being a n-degree polynomial in time, Hencic and Gouriéroux (2015), or the HP filter Hecq and Voisin (2021). Hecq and Voisin (2019) show the performance of these different available methods for estimation and forecasting performance. Unfortunately, these de-trending approaches often do not have an economic or financial interpretation. Moreover they are not able to deliver a meaningful forecast of the trend and bubble component. While the overall in-sample fit might be reasonable for high-order polynomials or HP-filters, the ability to produce reasonable forecasts are severely limited. Hecq and Voisin (2019) show in a simulation exercise that these methods may perform well in estimating the true parameters and the right order of the process. However, the use of these methods is problematic when a bubble is growing in real time. Indeed, the extreme movements observed during a locally explosive behaviour will always be captured as a trend by these detrending methods, and only ex-post recognized as a bubble, once the bubble collapses. This renders the current de-trending methods very problematic for forecasting in real-time.

In this paper we propose an observation driven model which jointly models the random-walk stochastic trend as well as the stationary non-causal bubble component of financial time-series. In line with the MAR literature, we assume an additive structure with a 'trend plus a bubble component'. However, in contrast with the current literature, we estimate the trend and bubble components jointly. We build on the approach followed by Blasques et al. (2022) for a non-stationary location model and

establish the asymptotic properties of the maximum likelihood estimator and provide sufficient conditions for strong consistency and asymptotic normality. Furthermore, we establish the stochastic properties of data generated by our model and show that our filter is capable of uncovering the unobserved stochastic trends and bubble components. In particular, the filters for the stochastic trend and stationary bubbles are shown to be invertible and to converge to the true trend and bubble paths. A Monte Carlo simulation reveals good finite sample properties for the estimator and the observation-driven filter of both the stochastic trend and the MAR bubble component. We finally consider an application on the NASDAQ composite index monthly series and compare the forecasting accuracy of our method against the detrending approaches used in the MAR literature. Our model is naturally capable of forecasting the trend component and the bubble component, where the latter rely on the methodology introduced in Gourieroux and Jasiak (2016) and Lanne and Saikkonen (2011). We show that our methods outperforms all the considered alternatives.

We revisit the application in Hecq and Voisin (2021) and find strong evidence for the existence of bubbles in the monthly Nickel price. In particular, we show that incorporating MAR dynamics substantially improves the model fit over random-walk and GARCH alternatives. Unlike Hecq and Voisin (2021) we do not need to de-trend the NASDAQ series with HP filters and can allow instead the fundamental value to follow a random walk, in line with financial theory, the efficient market hypothesis and countless papers in financial econometrics. We also provide a climate application using one of the time series considered by Giancaterini et al. (2022). Our model shows a good forecasting performance on sea level measurements, a non stationary time series that presents small but frequent non causal dynamics. In both applications we perform a real time forecasting exercise, showing how our models outperforms the alternatives.

This paper is organized as follows. In the next section we present our model specification. In section 3 we present the properties of the estimated model. In section 4 we perform an application using NASDAQ monthly price and in section 5 we have a simulation study to compare our model with the existing methods in the literature.

2 The Model

Consider a non-stationary time-series $\{y_t\}$ which can be decomposed into a non-stationary random-walk component $\{\mu_t\}$ and a stationary process $\{v_t\}$ with MAR(r, s) dynamics according to the following observation-driven model which we call a *MAR stochastic trend* model (MARST),

$$
\begin{aligned}
y_t &= \mu_t + v_t \\
v_t &= \psi(L^{-1})^{-1}\phi(L)^{-1}\varepsilon_t \\
\mu_t &= \delta + \mu_{t-1} + \alpha\varepsilon_{t-s}
\end{aligned} \tag{1}
$$

where $\psi(z) = 1 - \psi_1 z - ... - \psi_s z^s$ and $\phi(z) = 1 - \phi_1 z - ... - \phi_s z^s$ are respectively the lags and leads polynomial, $\{\varepsilon_t\}$ is an independent identically distributed sequence with $\varepsilon_t \sim t_\nu$ with $\nu > 0$ the degrees of freedom and σ a scale parameter. In the updating equation for μ_t, δ represents a possible drift in the trend component, α drives the amplitude of the update. We note that $k = r + s$ is the total order of the autoregressive polynomial and that the MARST model in (1) allows for an additive structure of the stationary bubble and non-stationary trend components.

The time-series $\{y_t\}$ will be rendered unit-root non-stationary, for any $\alpha > 0$, and feature a drift whenever $\delta \neq 0$, where $\{\mu_t\}$ can be interpreted as the fundamental value while $\{v_t\}$ captures the bubbles and the stationary autoregressive features through a MAR specification. The MAR component can be causal, non-causal, or both, depending on the parameters in the causal polynomial $\phi(z)$ or the non-causal polynomial $\psi(z)$. Assumption 1 imposes well known restrictions rendering the MAR(r, s) process $\{v_t\}$ stationary and ergodic; see e.g. Lanne and Saikkonen (2011).

Assumption 1 *The polynomials $\phi(z)$ and $\psi(z)$ satisfy,*

$$\phi(z) = 0 \quad for \quad |z| > 1 \quad and \quad \psi(z) = 0 \quad for \quad |z| > 1.$$

Naturally, the model defined in (1) nests a number of different important models available in the literature. For example, when (1) features a MAR$(0, 0)$ component, it defines a model with random walk dynamics (with or without drift), as in Blasques et al. (2022). If only the causal part of the autoregressive polynomial is non-zero, then the model contains short-run stationary causal dynamics but does not allow for bubbles. This means that by taking appropriate model selection steps, one can effectively consider a range of possibilities, and even if the model is designed to deal with bubbles in non-stationary time series it also allows to reject these features.

We also note that the MARST model in (1) can be re-written as a non-stationary MAR featuring appropriate parameter restrictions on the causal polynomial. In particular, the following remark highlights that the observation-driven specification with a non-stationary random-walk component plus a stationary MAR component is just one of multiple possible model representations.[1]

Remark 1 *Let Assumption 1 hold. Then $\{y_t\}$ satisfies the following mixed autoregressive integrated moving average, MARIMA(r,s,1,k+r), representation,*

$$\phi(L)\psi(L^{-1})\Delta y_t = \delta + \boldsymbol{\theta}(L)\varepsilon_t$$

[1]Derivations are made available in Appendix A.

where $\boldsymbol{\theta}(z) = 1 - z^s + \alpha_0 z^s \phi(z)\psi(z^{-1})$. *Further, if* $\psi(L^{-1}) = \phi(L) = 1$, *then model in (1) generates a random-walk process with drift with non-linear MA innovations* $y_t = \delta + y_{t-1} + \varepsilon_t + \gamma(\varepsilon_{t-1})$, *where* $\gamma(\varepsilon_{t-1}) = (\alpha - 1)\varepsilon_{t-1}$. *Moreover if* $\alpha = 1$ *then (1) defines a random-walk process* $y_t = y_{t-1} + \sigma\varepsilon_t$.

3 Estimation and filtering

Let the parameter vector be defined as $\boldsymbol{\theta} = (\delta, \alpha, \Psi, \boldsymbol{\gamma})$, with δ, α being parameters that drive the update of the filter in (1), $\Psi = (\phi_1, ..., \phi_r, \psi_1, ..., \psi_s)$ the vector of MAR parameters and with $\boldsymbol{\gamma}$ consisting of the vector of the distributional parameters $\boldsymbol{\gamma} = (\sigma, \nu)$. To estimate the parameter vector $\boldsymbol{\theta}$ we rely on the Approximate Maximum Likelihood framework for MAR processes from Lanne and Saikkonen (2011). The log-likelihood criterion function $L_T(\boldsymbol{\theta})$ is naturally given by,

$$\hat{L}_T(\boldsymbol{\theta}) = \sum_{t=r}^{T-s} \hat{l}_t(\boldsymbol{\theta}) = \sum_{t=r}^{T-s} \log f\Big(\psi(L^{-1})\phi(L)\hat{g}_t(\boldsymbol{\theta}); \boldsymbol{\gamma}\Big)$$

where l_t is the individual likelihood contribution, $f(\cdot)$ is the Student's t pdf and $\hat{g}_t(\boldsymbol{\theta}) = y_t - \hat{\mu}_t(\boldsymbol{\theta})$ is the bubble component. The hat symbol represents the dependence of these functions on the filtered bubble component $\hat{g}_t(\boldsymbol{\theta})$, instead of the limit counterpart $g_t(\boldsymbol{\theta})$. Moreover note that $\hat{g}_t(\boldsymbol{\theta})$ can also be interpreted as a prediction error, being the deviation of the process y_t from the filtered random walk component $\hat{\mu}_t(\boldsymbol{\theta})$. We are interested in the properties of maximum likelihood (ML) estimator defined as,

$$\hat{\boldsymbol{\theta}}_T \arg\max_{\boldsymbol{\theta}\in\Theta} \hat{L}_T(\boldsymbol{\theta}).$$

We note that the log-likelihood depends on the prediction error $\hat{g}_t(\boldsymbol{\theta})$ which must be obtained after the random walk component $\hat{\mu}_t(\boldsymbol{\theta})$ is 'filtered out'. In practice, for any given $\boldsymbol{\theta}$, and given a sample $\{y_t\}_{t=1}^T$, we obtain first the filtered sequence $\{\hat{\mu}_t(\boldsymbol{\theta})\}_{t=1}^T$,

$$\hat{\mu}_{t+1}(\boldsymbol{\theta}) = \omega + \hat{\mu}_t(\boldsymbol{\theta}) + \alpha\psi(L^{-1})\phi(L)\big(y_{t-s} - \hat{\mu}_{t-s}(\boldsymbol{\theta})\big)$$

and the initialization is given by the first k observations of the sample $(\hat{\mu}_k(\boldsymbol{\theta}), ..., \hat{\mu}_1(\boldsymbol{\theta}))$ $= (y_k, ..., y_1)$, where $k = r + s$ and r and s are respectively the causal and non-causal order of our MAR(r, s) process.

Given the non-stationary filter $\hat{\mu}_t(\boldsymbol{\theta})$, our interest lies in the stationarity of the prediction error $\hat{g}_t(\boldsymbol{\theta}) = y_t - \hat{\mu}_t(\boldsymbol{\theta})$. We can write the prediction error according to the following SRE,

$$\hat{g}_{t+1}(\boldsymbol{\theta}) = \hat{g}_t(\boldsymbol{\theta}) - \omega - \alpha\psi(L^{-1})\phi(L)\hat{g}_{t-s}(\boldsymbol{\theta}) + \Delta y_{t+1}.$$

Further, we can define the vector $\hat{\mathbf{g}}_t = \left[\hat{g}_t, \hat{g}_{t-1}, ..., \hat{g}_{t-k+1}\right]'$ and write the model residuals as $\hat{\varepsilon}_{t-s}(\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta})'\hat{\mathbf{g}}_t(\boldsymbol{\theta})$, and their limit counterpart as $\varepsilon_{t-s}(\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta})'\mathbf{g}_t(\boldsymbol{\theta})$, with $\Phi(\boldsymbol{\theta}) = (\zeta_0, ..., \zeta_{k-1})$ and

$$\begin{cases} \zeta_i &= \sum_{j=1}^{i} \psi_{s-j+1}\phi_{i-j} \quad \text{for } i = 1, ..., s \\ \\ \zeta_i &= \sum_{j=1}^{k-i+1} \psi_{j-1}\phi_{j+i-s-1} \quad \text{for } i = s+1, ..., k. \end{cases}$$

Neither should be confused with ε_t the iid noise of (1). From now on we will drop the subscript $-s$ for the residuals when we work with the vector form of the prediction errors and their limit counterpart, respectively $\hat{\mathbf{g}}_t(\boldsymbol{\theta}), \mathbf{g}_t(\boldsymbol{\theta})$ while it should be kept in mind that any $\varepsilon_t(\boldsymbol{\theta})$ depends on the items $g_{t-r}, ..., g_{t+s}$. Then, we can write our SRE in vector form as,

$$\hat{\mathbf{g}}_{t+1}(\boldsymbol{\theta}) = C(\boldsymbol{\theta}) + A(\boldsymbol{\theta})\hat{\mathbf{g}}_t(\boldsymbol{\theta}) + B_{t+1},$$

with

$$C(\boldsymbol{\theta}) = \begin{bmatrix} -\delta \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad A(\boldsymbol{\theta}) = \begin{bmatrix} 1-\alpha\zeta_1 & -\alpha\zeta_2 & \dots & -\alpha\zeta_{k-1} & -\alpha\zeta_k \\ 1 & 0 & \dots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \quad B_{t+1} = \begin{bmatrix} \Delta y_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where, to keep notation short, we will often redefine $A(\boldsymbol{\theta})$ as,

$$A(\boldsymbol{\theta}) = \begin{bmatrix} \xi_1 & \xi_2 & \dots & \xi_{k-1} & \xi_k \\ 1 & 0 & \dots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}$$

Stationarity and filter invertibility

The stationarity of the MAR component of the data and the filter invertibility are important ingredients in establishing the consistency of the MLE, as they allow for laws of large numbers to be applied to the log likelihood loss function. In order to establish the stationarity of the data generating process of the underlying MAR process and the invertibility of the filter, we impose a number of restrictions on the parameter space. Specifically, we assume that the parameters of the MAR define stable or contracting process, and that the innovations have n moments.

Assumption 2 *The degrees of freedom parameter for the Student's t innovations ν satisfies $\nu \geq n > 1$, so that $\mathbb{E}\left|\varepsilon_t\right|^n < \infty$ for some $n > 1$.*

Assumption 3 *The parameters in the matrix $A(\boldsymbol{\theta})$ are such that $\sup_{\boldsymbol{\theta}\in\Theta} |\xi_1| < 1$ and $\sup_{\boldsymbol{\theta}\in\Theta} \left|\sum_{i=1}^{k} \xi_i\right| < 1$.*

Proposition 1 shows the properties of data generated by the model in (1).

Proposition 1 *Let $\{y_t\}_{t\in\mathbb{Z}}$ be generated by the model defined in (1) with some true parameter vector $\boldsymbol{\theta}_0 \in \boldsymbol{\theta}$. Under Assumptions 1 and 2 we have that $\{\Delta y_t\}_{t\in\mathbb{Z}}$ is a stationary and ergodic process with $\mathbb{E}|\Delta y_t|^n < \infty$.*

Proposition 2 gives sufficient conditions for the uniform invertibility of the filter $\hat{\mathbf{g}}_t(\boldsymbol{\theta})$, establishing its convergence to a unique stationary and ergodic process, as well as the convergence of the corresponding model residuals.

Proposition 2 *Let $\{y_t\}_{t\in\mathbb{Z}}$ be generated by the model defined in (1). Let Assumptions 1-3 hold. Then we have that,*

a) *the filtered sequence of prediction errors satisfies,*

$$\sup_{\boldsymbol{\theta}\in\Theta} ||\hat{\mathbf{g}}_t(\boldsymbol{\theta}) - \mathbf{g}_t(\boldsymbol{\theta})|| \xrightarrow{e.a.s.} 0, \quad \text{as } t \to \infty.$$

with $\{\mathbf{g}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ a unique, stationary and ergodic sequence.

b) *the residual $\hat{\varepsilon}_t(\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta})\hat{\mathbf{g}}_t(\boldsymbol{\theta})$ and its limit counterpart $\varepsilon_t(\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta})\mathbf{g}_t(\boldsymbol{\theta})$ satisfy,*

$$\sup_{\boldsymbol{\theta}\in\Theta} ||\hat{\varepsilon}_t(\boldsymbol{\theta}) - \varepsilon_t(\boldsymbol{\theta})|| \xrightarrow{e.a.s.} 0, \quad \text{as } t \to \infty.$$

Propositions 3 and 4 show that the stationary limit solutions identified in Proposition 2 have n bounded moments, and that the filter $\{\hat{\mu}_t(\boldsymbol{\theta}\}$ of the non-stationary component converges to the true $\{\mu_t\}$.

Proposition 3 *Under Assumptions 1-3, the limit process $\{\mathbf{g}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ is such that $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} ||\mathbf{g}_t(\boldsymbol{\theta})||^n < \infty$. Moreover $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} ||\varepsilon_t(\boldsymbol{\theta})||^n < \infty$.*

Consistency and Asymptotic Normality

We now establish the strong consistency of the MLE $\hat{\boldsymbol{\theta}}_T$. We make use of the compactness of the parameter space.

Assumption 4 *Θ is a compact set such that (1) holds for every $\boldsymbol{\theta} \in \Theta$.*

Theorem 1 *Under Assumptions 1,3-4, and if $\mathbb{E}|\varepsilon_t|^n < \infty$ for $n > 2$ then MLE $\hat{\boldsymbol{\theta}}_T$ satisfies $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$ as $T \to \infty$.*

The proof follows the approach of Blasques et al. (2018) and Straumann and Mikosch (2006). Given the filter invertibility and the moment conditions established in the previous section, asymptotic consistency follows by following the same concepts for the approximate maximum likelihood consistency in Lanne and Saikkonen (2011)

and Breid et al. (1991) accounting for the fact that our setting includes a set of additional parameters.
We finally turn to the $\sqrt{T}$-convergence and asymptotic normality of the MLE. We first establish relevant properties for the derivatives of the limit filter. These properties will play an important role in showing that the estimator is asymptotically normal.

Proposition 4 *Let $\{y_t\}_{t\in\mathbb{Z}}$ be generated by (1). Under Assumptions 1-3 we have,*

(a) $\{\partial\hat{\mathbf{g}}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}_{t\in N}$ converges e.a.s. to a unique stationary and ergodic sequence $\{\partial\mathbf{g}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}_{t\in\mathbb{Z}}$ uniformly over $\boldsymbol{\theta}$ such that $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}||\partial\mathbf{g}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}||^n<\infty$;

(b) $\{\partial^2\hat{\mathbf{g}}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\}_{t\in N}$ converges e.a.s. to a unique stationary and ergodic sequence $\{\partial^2\mathbf{g}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\}_{t\in\mathbb{Z}}$ uniformly over $\boldsymbol{\theta}$ such that $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}||\partial^2\mathbf{g}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'||^n<\infty$.

To obtain the asymptotic normality of the MLE we need to assume the existence of additional moments. The following assumption requires the Student's t distributed innovations to have more than four degrees of freedom.

Assumption 5 *The degrees of freedom parameter for the Student's t innovations ν satisfies $\nu\geq n>4$, so that $\mathbb{E}\left|\varepsilon_t\right|^n<\infty$ for some $n>4$.*

Theorem 2 *Assume that assumptions 1,3-5 hold. Let $\boldsymbol{\theta}_0$ lie in the interior of Θ. Then we have that*

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta})\xrightarrow{d}\mathcal{N}(0,\mathcal{I}^{-1})\ \text{as}\ T\to\infty$$

where $\mathcal{I}=-\mathbb{E}[\partial^2 l_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']$ is the Fisher information matrix.

For this theorem we follow the argument of Theorem 3.1 of Gorgi and Koopman (2021) and Section 7 of Straumann and Mikosch (2006). The additional moments on the innovations are required in order to ensure the existence of the variance of the estimator.

4 Monte Carlo study

This section uses a Monte Carlo simulation exercise to show that (i) the MARST model can effectively filter the stochastic trend and uncover the short-term explosive bubble behavior in a simulated dataset; and (ii) to analyze the distortionary effect of different detrending techniques on the estimation of bubble dynamics as in Hecq and Voisin (2019).

Filtering trends and bubbles

In general, most financial time series are not well described by a pure MAR process, where bubbles are pervasive and generated continuously. In contrast, it seems that financial time series exhibit a small number of bubbles, often characterized by a great magnitude. To mimic this feature we simulated a number of random walks and added occasional large bubbles generated by a MAR process with infrequent but very large errors. The sum of these two processes creates a random walk with only a few locally explosive episodes. We find that our model is able to disentangle the two processes even if the MAR component is not always present, in other words, the model model allows the non-causal part to disappear in moments where there is no-bubble and to be activated when there is a sudden bubble.

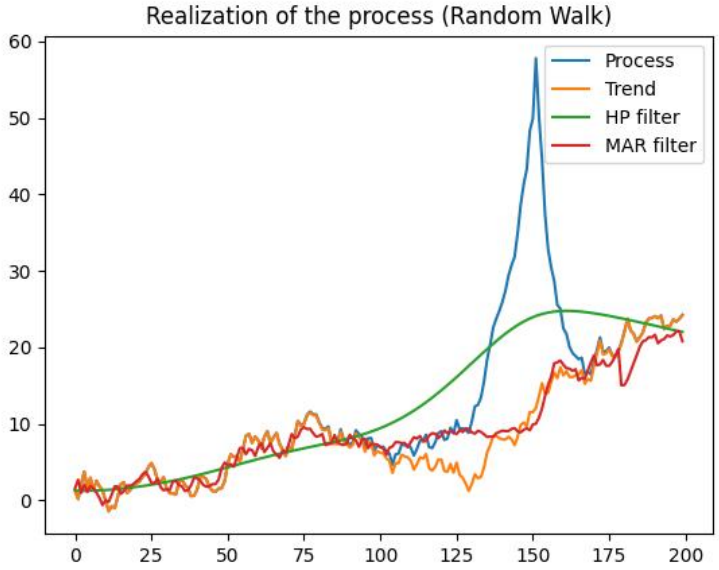


Figure 1: The trend obtained through the MAR + trend filter follows closely the random walk until the bubble component kicks in. The HP filter instead has problems to detect what is the trend component and what is not.

Figure 1 shows an example of such simulated path and the filtered stochastic trend obtained by both our model and the HP filter.

Detrending and MLE distortion

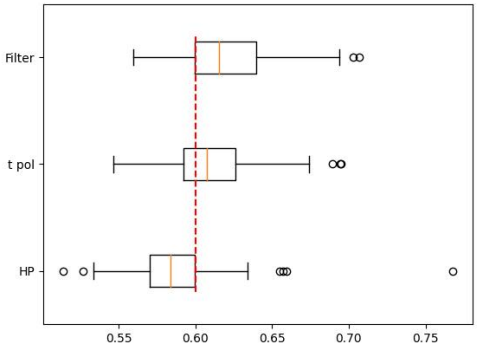
The following Monte Carlo simulation follows the procedure from Hecq and Voisin (2019) in documenting how different de-trending procedures distort parameter estimates in a range of data generating processes. In particular, we simulate data from a number of mispecified sources and find that the current model can easily identify

the non-causal process in addition to a given trend. We also simulate a MAR process without trend to show that the stationary mixed causal non-causal process (which is a special case of our model) does not result in distorted ML parameter estimates. Specifically, we consider the following data generating processes[2]: (a) simple MAR process; (b) MAR process plus a random walk with drift; (b) MAR process plus deterministic trend breaks.

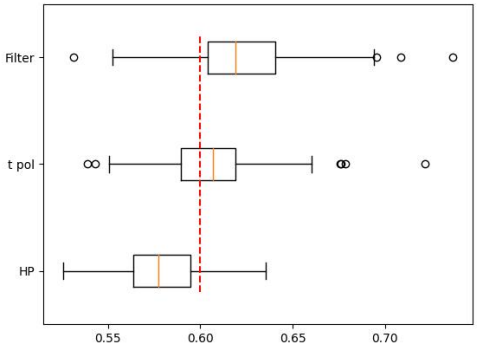
The sample without a trend is used to analyze distortions on the MLE when a trend is actually not present. We use the random walk with drift to represent the type of composition of stochastic trend and bubble process that would be usually observable in real data. The last data generating process we consider represents a situation with breaks in the trend to mimic the process defined in the simulation study from Hecq and Voisin (2019).

We simulate $S = 100$ samples of length $T = 400$ of MAR$(1, 1)$ processes with $\psi = \phi = 0.6$ and degrees of freedom $\nu = 2$. In Figure 2 we present the distribution of the estimates obtained according to the different estimation methods that we use, respectively our MARST model and the time polynomial or HP filter detrending plus MAR. In the figure the ϕ and the ψ estimates are presented in sequence for all the considered data generating processes. As we can see the type of distortion produced in the parameter estimates by our model is similar to the one observed using the standard methods in the literature.

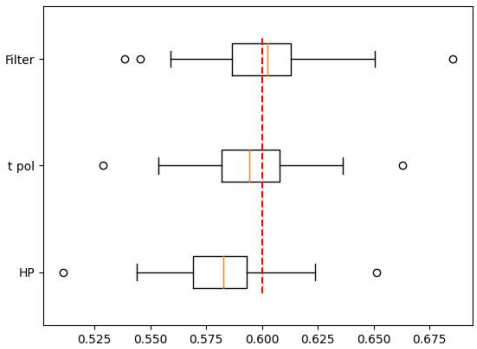
[2]Figure C in the Appendix shows examples of realized paths from such data generating processes.



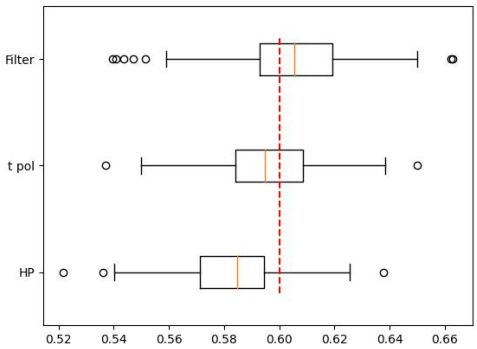
(a) ψ estimates for a random walk with drift dgp.



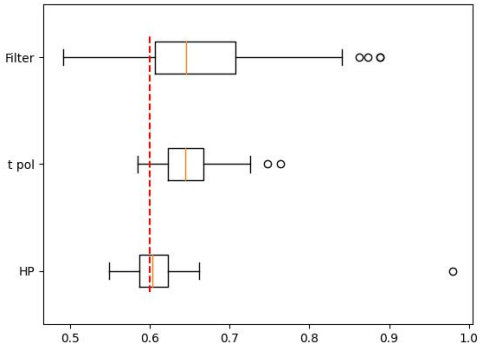
(b) ϕ estimates for a random walk with drift dgp.



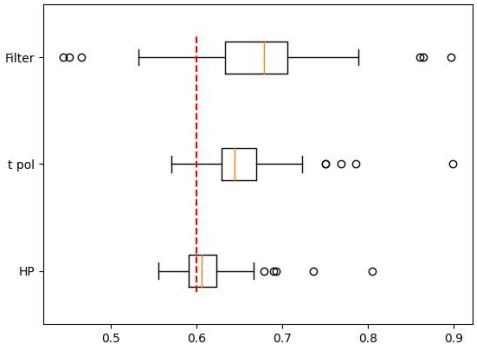
(c) ψ estimates for no trend dgp.



(d) ϕ estimates for no trend dgp.



(e) ψ estimates for a dgp with breaks in the trend.



(f) ϕ estimates for a dgp with breaks in the trend.

Figure 2: Distributions of the monte carlo estimates of the causal and non causal parameters

5 Empirical Applications

In this section we go through two different applications. The first one is relevant in commodity pricing and considers the monthly price of Nickel which is characterized by very sizeable bubbles, especially in the period between 2007 and 2008. We show how our model outperforms the alternatives in terms of forecasting power. We also highlight how the model performance behaves over the different phases that the process experiences, and how our model is flexible enough to adapt to these phases. The second application focuses on data which is relevant in climate studies and considers a time series of global mean sea levels. This data is characterized by a strong trend and many smaller episodes of non causal dynamics. We find that modelling these periods of short-lived explosive behvior can substantially improve the ability of the model to fit the data.

5.1 Bubbles in commodity prices

We follow Hecq and Voisin (2021) in analyzing the seemingly large bubble present in the monthly time-series of global Nickel Price. Our interest lies in understanding if the presence of a MAR component introduced by the MARST model beyond the usual random-walk will lead to an improvement of the online forecast of the nickel price. Figure 3 plots the Nickel price spanning from January 1990 to October 2022.

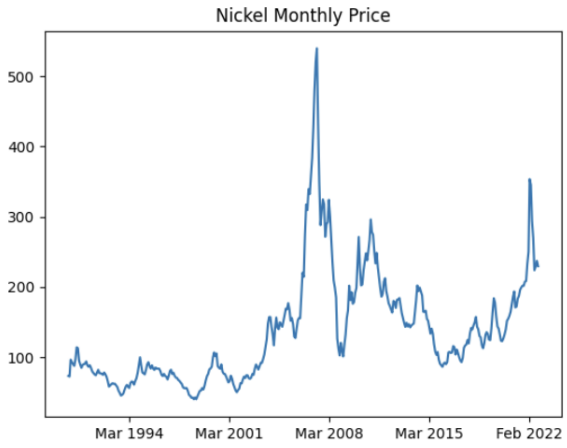


Figure 3: Monthly Nickel Price

A natural consequence of the MARST model is the ability to extract both a stochastic trend and a MAR component. In light of the theory developed in Section 3, and the Monte Carlo evidence in Section 4, the divergence of the MAR component from the stochastic trend should give us an indication of whether locally explosive bubble dynamics are present or not.

For the model selection step we consider a MARST(r, s) with $r \leq 5$ and $s \leq 5$. We choose the causal and non-causal model order according to BIC, that indicates a MAR$(1, 1)$ as the most suitable option.

Order	Likelihood	AIC	BIC
(1,1)	604.81	1221.63	**1241.42**
(2,3)	602.10	1222.20	1251.88
(4,2)	597.99	**1215.98**	1248.97
(4,3)	**597.46**	1216.92	1253.20
(4,4)	598.10	1220.21	1259.79

Table 1: Model Selection Criteria for the Nickel Monthly price application.

We make use of interval forecasts to show that our model manages to predict the sign of the extreme observations during a bubble in a consistent way. We report our performance compared to the performance of a random walk model and a random-walk with GARCH volatility as the existing versions of online forecasting using MAR models have very poor performances in non stationary settings due to their lack of a reliable method to forecast the trend part.

Beyond comparing interval forecasts, we further perform a point forecast. While we report on the usual point forecast, corresponding to the conditional expectation, we shall pay closer attention to the interval forecast. Point forecasts in the MAR framework are less suitable for interpretation and less informative than in most other contexts. This is because in the MAR framework predictive density is bimodal whenever there is am active bubble component. This bimodality comes from the bubble behavior which attaches a given probability to the event that the bubble continues, and some probability to the event that the bubble will crash; see e.g. Hecq and Voisin (2021). [3]

Our test sample includes the last 190 observations of the sample, spanning from September 2006 to October 2022. We select this test sample to show two main advantages of our model. The data is shown in Figure 4. First, it is able to provide reliable forecasts during the 2008 bubble, that is the most relevant in the sample. Second our approach performs well also when the non-stationarity of the data becomes more evident, so from 2009 onwards.

[3]With the exception of the MAR(r,1) with Cauchy innovations, see Gouriéroux and Zakoïan (2013), there is no closed form solution for the predictive density of most MAR specifications. Instead, there are two approximation methods to compute forecasts in this framework. One that is simulation based (Lanne et al. (2012)), and another which is sample based (Gourieroux and Jasiak (2016)). We rely on the method from Lanne et al. (2012).

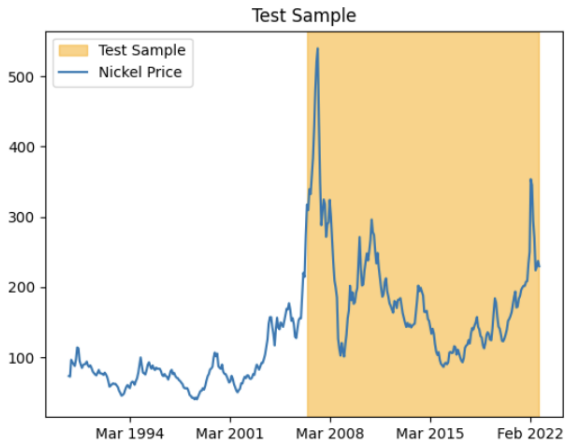


Figure 4: The time series of Nickel price. The test sample is highlighted in orange.

We work with an expanding window and with the simulation based forecast method of Lanne et al. (2012). Since the method is based on simulation, the forecast is subject to some variability, regardless Hecq and Voisin (2021) show that the approximation behaves coherently with the theoretical probabilities. We set the number of iterations to a high threshold to reduce the uncertainty to a minimum. To compare the performance of the different methods we rely on a Diebold Mariano test statistic based on the Brier score which is a well suited performance measure for discrete outcomes [4]. Table 2 summarizes the results.

Models	Brier Score	
	Model Score	Test Statistic
MARST	**0.37**	·
MAR (a)	0.45	**-1.67**
Random Walk (b)	0.33	1.05
Random Walk (c)	0.44	**-2.04**

Table 2: Event Prediction Scores and Test Statistics against **a)** MAR with time polynomial, **b)** Random Walk with GARCH specification and Student's t innovations, **c)** Gaussian Random Walk.

We now show how our method behaves compared to the alternatives in different parts of the sample. We can split our large test sample in three subsamples: a subsample for the 2008 bubble, a huge non-causal event but with a low level of non stationarity, a subsample from 2009 to 2018, exihibiting a low level of non-causality, being very close to a simple random walk, and a subsample for the last part of the sample where again we non causal events and a seemingly relevant upwards trend. Figure 5 summarizes the division in subsamples.

[4]See Appendix D for additional details on the testing procedure

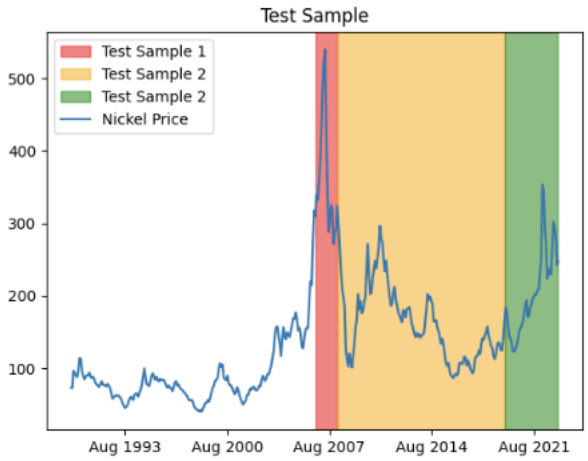


Figure 5: Testing subsamples subdivision of the whole sample

In the first subsample the bubble component is dominant, the random walk with Gaussian innovations fails to capture the extreme movements that the process is experiencing. During this period the robust random walk performs a bit better but we still outperform both methods. In the second subsample the process experiences a path that can be approximated well by a random walk. Our model still outperforms the Gaussian Random Walk, but has a worse performance than the more flexible GARCH driven process, however the difference are not significant. The intuition behind the better performance of the random walk with Garch innovations compared to the MARST comes from the fact that in the considered subsample, without relevant non-causal episodes the MARST dynamics mimics the one of a simple random walk. In the third subsample the process experiences again relevant explosive episodes. For this reason the random walk with Gaussian innovations does not manage to capture the movements of the process during the small but sharp bubbles we can see in this part of the sample. Table 3 shows the test statistics over these three different subsamples.

	$1^{st}Subsample$		$2^{nd}Subsample$		$3^{rd}Subsample$	
Models	Brier Score		Brier Score		Brier Score	
	Score	Statistic	Score	Statistic	Score	Statistic
MARST	**0.44**	·	0.45	·	**0.24**	·
RW (a)	1.20	**-13.88**	0.44	0.19	0.32	**-2.16**
RW (b)	0.60	**-2.36**	**0.38**	1.26	**0.24**	0.59

Table 3: Event Prediction Scores and Test Statistics for the Nickel Price application over the three different test subsamples **a)** Gaussian Random Walk , **b)** Random Walk with Garch specification and Student's t innovations.

5.2 Modeling sea levels

In this second application we investigate the presence on non-causal dynamics in global mean sea level data. As we shall se the MARST model manages to capture a stochastic trend without relying on methods like HP filter. Further we will note that the MARST provides us with reliable forecasts.
Climate change has grown to become of the most debated challenges of our times. Understanding the dynamics of climate data is a crucial step in the debate. Giancaterini et al. (2022) use MAR models to assess the time reversibility of different climate change indicators like emission of greenhouse gases, temperature anomalies or global mean sea level. These time series show clear positive trends and a certain degree of non-causality, so they are good candidates for the current dynamic model.

Figure 6 plots the global mean sea-level data. The data spans from 2011 to 2022. We use the last 30 observations of the sample as pseudo-out-of-sample observations to test the performance of the one step ahead forecasts from our model.

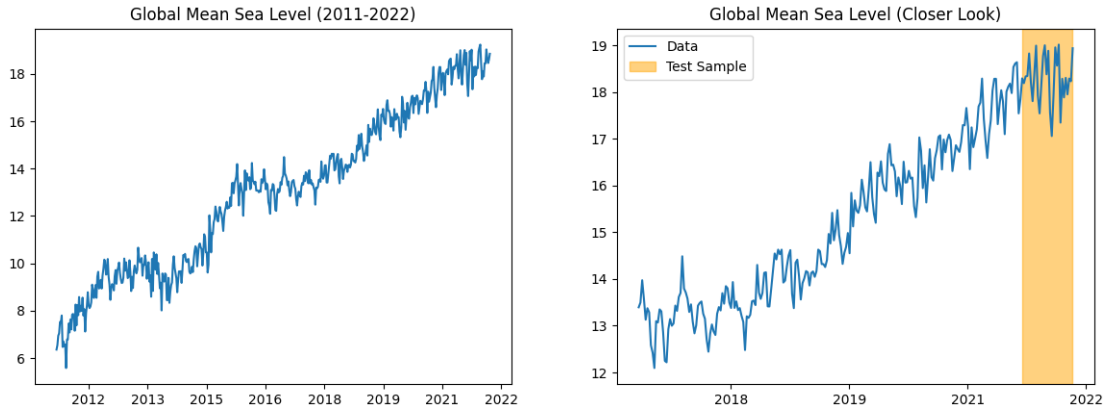


Figure 6: Global mean sea level time series (left), and the most recent period (right) with a highlight of the period we will use as our test sample.

For the model selection step we consider a MARST(r, s) model with $r \leq 5$ and $s \leq 5$. Table 4 shows standard information criteria at different lag/lead lengths. We select the model order following the BIC and adopt a MARST(1,1) specification.

Figure 6 presents the fitted MARST trend and the resulting detrended MAR part. We observe that the MARST model captures well the stochastic trend component of the process.

	Likelihood	AIC	BIC
(1,1)	271.13	554.26	**579.55**
(2,1)	271.66	557.32	586.82
(2,2)	273.10	562.20	595.92
(3,4)	260.05	**542.10**	588.46
(4,4)	**259.99**	543.98	594.55

Table 4: Model Selection Criteria for the Global Mean sea level application

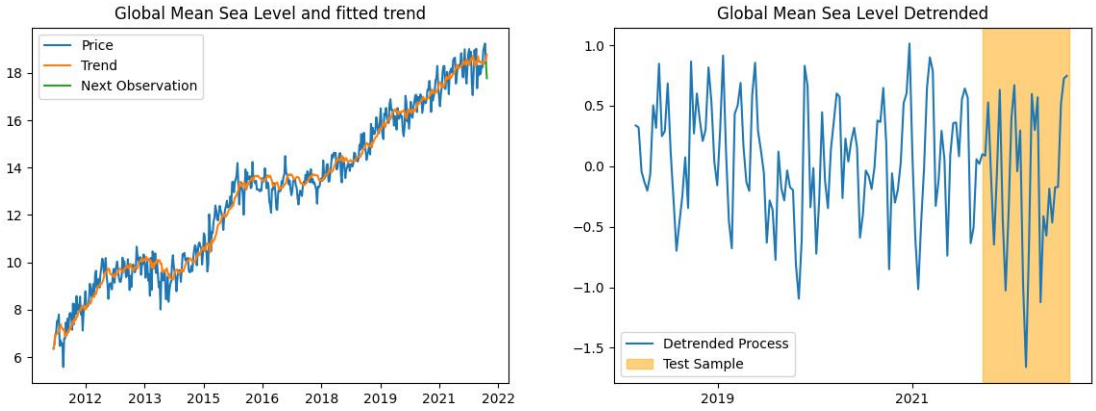


Figure 7: Global mean sea level time series with the trend fitted by our model (left),detrended time series, in other words the remainder MAR process (right). We zoomed in on the last 150 observations to have a closer look at the small bubbles appearing in the sample.

We compare our forecasting performance against a MAR process that relies on a standard detrending procedure, in this case a time polinomyal, and against a random walk process, both Gaussian and heavy tailed. We also compare our model against an ARIMA specification. Recall from Remark 1 that some ARIMA and MARIMA specifications are nested into our model. We still expect the ARIMA specification to perform reasonably well in parts of the sample where the non causal part is not dominant. We still rely on interval forecast performance as the bimodality of the predictive density renders point forecast accuracy a rather uninteresting performance measure for comparing models with such rich predictive behavior. Table 5 shows that the MARST model outperforms all competing models according to a Diebold Mariano test statistic based on the Brier Score [5].

[5]see Section D in Appendix for additional details on the testing procedure

Models	Brier Score	
	Model Score	Test Statistic
MAR with trend	**0.52**	·
MAR (a)	0.63	**-2.83**
ARIMA (2,1,2) (b)	0.59	**-1.99**
Random Walk (c)	0.66	**-2.33**

Table 5: Event Prediction Scores and Test Statistics for the sea level application against **a)** MAR with time polynomial, **b)** ARIMA specification **c)** Random Walk with Gaussian innovations.

6 Conclusion

We proposed a new dynamic model which can jointly filter stochastic trends and stationary MAR components from time-series data. The model was shown to be relevant for handling non-stationary time series with random-walk trends and locally explosive behavior. In comparison with existing de-trending approaches used in the MAR literature, our method has the advantage of allowing for online forecast of financial bubbles in non-stationary time series. We showed in the empirical application data that our method allows us to forecast the data during bubble episodes better than using the existing methods.

References

Blanchard, O. J. and Watson, M. W. (1982). Bubbles, rational expectations and financial markets. *NBER working paper*, w0945.

Blasques, F., Gorgi, P., Koopman, S. J., and Wintenberger, O. (2018). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics*, 12(1):1019 – 1052.

Blasques, F., van Brummelen, J., Gorgi, P., and Koopman, S. (2022). Maximum likelihood estimation for non-stationary location models with mixture of normal distributions. 22-001.

Breid, F., Davis, R. A., Lh, K.-S., and Rosenblatt, M. (1991). Maximum likelihood estimation for noncausal autoregressive processes. *Journal of Multivariate Analysis*, 36(2):175–198.

Chen, X. and Funke, M. (2013). Real-time warning signs of emerging and collapsing chinese house price bubbles. *National Institute Economic Review*, 223(1):R39–R48.

Diba, B. T. and Grossman, H. I. (1988). Explosive rational bubbles in stock prices? *The American Economic Review*, 78(3):520–530.

Etienne, X. L., Irwin, S. H., and Garcia, P. (2014). Bubbles in food commodity markets: Four decades of evidence. *Journal of International Money and Finance*, 42:129–155.

Evans, G. W. (1991). Pitfalls in testing for explosive bubbles in asset prices. *The American Economic Review*, 81(4):922–930.

Fries, S. and Zakoïan, J.-M. (2017). Mixed causal-noncausal ar processes and the modelling of explosive bubbles. *CREST working paper*.

Giancaterini, F., Hecq, A., and Morana, C. (2022). Is climate change time-reversible? *Econometrics*, 10:36.

Gorgi, P. and Koopman, S. (2021). Beta observation-driven models with exogenous regressors: A joint analysis of realized correlation and leverage effects. *Journal of Econometrics*, page 105177.

Gourieroux, C. and Jasiak, J. (2016). Filtering, prediction and simulation methods for noncausal processes. *Journal of Time Series Analysis*, 37(3):405–430.

Gouriéroux, C. and Zakoïan, J.-M. (2013). Explosive bubble modelling by noncausal process. *CREST working paper*.

Gouriéroux, C. and Zakoïan, J.-M. (2017). Local explosion modelling by non-causal process. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):737–756.

Gutierrez, L. (2013). Speculative bubbles in agricultural commodity markets. *European Review of Agricultural Economics*, 40(2):217–238.

Hecq, A. and Voisin, E. (2019). Predicting crashes in oil prices during the covid-19 pandemic with mixed causal-noncausal models.

Hecq, A. and Voisin, E. (2021). Forecasting bubbles with mixed causal-noncausal autoregressive models. *Econometrics and Statistics*, 20:29–45.

Hencic, A. and Gouriéroux, C. (2015). Noncausal autoregressive model in application to Bitcoin/USD exchange rates. In *Econometrics of Risk*, pages 17–40. Springer, New York.

Homm, U. and Breitung, J. (2012). Testing for speculative bubbles in stock markets: a comparison of alternative methods. *Journal of Financial Econometrics*, 10(1):198–231.

Lanne, M., Luoto, J., and Saikkonen, P. (2012). Optimal forecasting of noncausal autoregressive time series. *International Journal of Forecasting*, 28(3):623–631.

Lanne, M. and Saikkonen, P. (2011). Noncausal autoregressions for economic time series. *Journal of Time Series Econometrics*, 3(3):1–32.

Milunovich, G., Shi, S., and Tan, D. (2019). Bubble detection and sector trading in real time. *Quantitative Finance*, 19(2):247–263.

Phillips, P. C. and Shi, S. (2019). Detecting financial collapse and ballooning sovereign risk. *Oxford Bulletin of Economics and Statistics*, 81(6):1336–1361.

Phillips, P. C., Shi, S., and Yu, J. (2015). Testing for multiple bubbles: Historical episodes of exuberance and collapse in the s&p 500. *International economic review*, 56(4):1043–1078.

Phillips, P. C. and Shi, S. P. (2018). Financial bubble implosion and reverse regression. *Econometric Theory*, 34(4):705–753.

Phillips, P. C., Wu, Y., and Yu, J. (2011). Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values? *International Economic Review*, 52(1):201–226.

Phillips, P. C. and Yu, J. (2011). Dating the timeline of financial bubbles during the subprime crisis. *Quantitative Economics*, 2(3):455–491.

Shi, S. (2017). Speculative bubbles or market fundamentals? an investigation of us regional housing markets. *Economic Modelling*, 66:101–111.

Straumann, D. and Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach. *The Annals of Statistics*, 34(5):2449–2495.

West, K. D. (1987). A specification test for speculative bubbles. *The Quarterly Journal of Economics*, 102(3):553–580.

Yiu, M. S. and Jin, L. (2012). Detecting bubbles in the hong kong residential property market: An explosive-pattern approach. *Hong Kong Institute for Monetary Research Working Paper*, (1).

A Proofs

A.1 Proof of Remark 1

Proof Consider a MAR(1, 1) with trend process. Then $\phi(L)\psi(L^{-1})(y_t - \mu_t) = \varepsilon_t$ with $\mu_t = \delta + \mu_{t-1} + \alpha_0\varepsilon_{t-s}$. As a result,

$$\phi(L)\psi(L^{-1})(y_t - \delta - \mu_{t-1} - \alpha_0\varepsilon_{t-s}) = \varepsilon_t.$$

We can now express $\mu_{t-1} = y_{t-1} - \big(\phi(L)\psi(L^{-1})\big)^{-1}\varepsilon_{t-s}$ and obtain the first claim,

$$\begin{aligned} & \phi(L)\psi(L^{-1})\Delta y_t = \omega^* - \varepsilon_{t-s} + \alpha_0\phi(L)\psi(L^{-1})\varepsilon_{t-s} + \varepsilon_t. \\ \Leftrightarrow \quad & \phi(L)\psi(L^{-1})\Delta y_t = \omega^* + \boldsymbol{\theta}(L)\varepsilon_t. \end{aligned}$$

For the second claim, we observe that,

$$\Delta y_t = \Delta\mu_t + \Delta\varepsilon_t \quad \Leftrightarrow \quad \Delta y_t = \delta + \alpha\varepsilon_{t-1} + \Delta\varepsilon_t.$$

Finally, if $\delta = 0$ and $\alpha = 1$, then it follows that $\Delta y_t = \varepsilon_t$. □

A.2 Proof of Proposition 1

Proof: By letting $y_t = v_t - \mu_t$, we can express

$$\begin{aligned} \Delta y_t &= \Delta\mu_t + \Delta v_t \\ \Delta y_t &= \delta_0 + \alpha_0\varepsilon_{t-k} + \Delta v_t \end{aligned}$$

where v_t is a MAR process that is SE under Assumption 1. Then we have that $\{\Delta y_t\}$ is SE as it is a measurable function of a SE process (Krengel, 1985, Proposition 4.3). Moreover, we have that

$$\mathbb{E}|\Delta y_t|^n \leq c_0 + c_1\mathbb{E}|\varepsilon_{t-k}|^n + c_2\mathbb{E}|\Delta v_t|^n < \infty$$

by the moment bound in Assumption 2. □

A.3 Proof of Proposition 2

Proof: We first note that under Assumptions 1 and 2, the spectral radius of $A(\boldsymbol{\theta})$ is smaller than one over all $\boldsymbol{\theta}$. This means that we have,

$$\sup_{\boldsymbol{\theta}\in\Theta} \big|\big|A(\boldsymbol{\theta})^r\big|\big| \leq K\rho^r \tag{2}$$

with $\rho < 1$. To show that **a)** holds, we define the infinite sum process,

$$\mathbf{g}_{t+1}(\boldsymbol{\theta}) = \sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r C(\boldsymbol{\theta}) + \sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r B_{t+1-r}. \tag{3}$$

So $\mathbf{g}_{t+1}(\boldsymbol{\theta})$ can be represented as the infinite sum of elements of the sequence $\{\Delta y_{t+1}\}$. By (2) we have that these sums converge. Since by Proposition 1 $\{\Delta y_t\}$ is a SE sequence, and since $\mathbf{g}_t$ is a continuous function of $\{\Delta y_t\}$ for every $\boldsymbol{\theta} \in \Theta$, we have that $\{\mathbf{g}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ is stationary and ergodic (Proposition 4.3, Krengel, 1985). Alternatively this process can also be written as the unique stationary solution of a vector AR(k) process according to Bougerol and Picard (1992).

Let's now consider the filter,

$$\hat{\mathbf{g}}_{t+1}(\boldsymbol{\theta}) = C(\boldsymbol{\theta}) + A(\boldsymbol{\theta})\hat{\mathbf{g}}_t(\boldsymbol{\theta}) + B_{t+1}$$

For a given initialization $\boldsymbol{\mu}_k = [\hat{\mu}_k, ..., \hat{\mu}_1]$ the corresponding $\hat{\mathbf{g}}_k = \left[\hat{v}_k, \hat{v}_{k-1}, ..., \hat{v}_1\right]'$ is a vector of fixed values and $\|\hat{\mathbf{g}}_k\|$ is finite. We can now unfold the process.

$$\hat{\mathbf{g}}_{t+1}(\boldsymbol{\theta}) = \sum_{r=0}^{t-k} A(\boldsymbol{\theta})^r C(\boldsymbol{\theta}) + A(\boldsymbol{\theta})^{t-k}\hat{\mathbf{g}}_k(\boldsymbol{\theta}) + \sum_{r=0}^{t-k} A(\boldsymbol{\theta})^r B_{t+1}, \tag{4}$$

and take the difference from the limit process,

$$\begin{aligned}
\|\mathbf{g}_{t+1}(\boldsymbol{\theta}) - \hat{\mathbf{g}}_{t+1}(\boldsymbol{\theta})\|^{\boldsymbol{\theta}} &= \|\sum_{r=t-k+1}^{\infty} A(\boldsymbol{\theta})^r C(\boldsymbol{\theta}) - A(\boldsymbol{\theta})^{t-h}\hat{\mathbf{g}}_k + \sum_{r=t-k+1}^{\infty} A(\boldsymbol{\theta})^r B_{t+1-r}\|^{\boldsymbol{\theta}} \\
&\leq \|A(\boldsymbol{\theta})^{t-k}\|^{\boldsymbol{\theta}} \cdot \|\sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r C(\boldsymbol{\theta}) + \sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r B_{k+1-r}||^{\boldsymbol{\theta}} + \|A(\boldsymbol{\theta})^{t-k}\|^{\boldsymbol{\theta}} \cdot \|\hat{\mathbf{g}}_k\| \\
&\leq K\rho^{t-k}\left(\|\sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r C(\boldsymbol{\theta}) + \sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r B_{k+1-r}\|^{\boldsymbol{\theta}} + \|\hat{\mathbf{g}}_k\|\right),
\end{aligned}$$

which implies that $\|\hat{\mathbf{g}}_t(\boldsymbol{\theta}) - \mathbf{g}_t(\boldsymbol{\theta})\|^{\boldsymbol{\theta}} \xrightarrow{e.a.s} 0 \quad$ as $T \to \infty$. To prove **b)** we note that,

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\Theta} ||\hat{\varepsilon}_t(\boldsymbol{\theta}) - \varepsilon_t(\boldsymbol{\theta})|| &= \sup_{\boldsymbol{\theta}\in\Theta} ||\Phi(\boldsymbol{\theta})\hat{\mathbf{g}}_t(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\mathbf{g}_t(\boldsymbol{\theta})|| \\
&\leq \sup_{\boldsymbol{\theta}\in\Theta} ||\Phi(\boldsymbol{\theta})|| \cdot \sup_{\boldsymbol{\theta}\in\Theta} ||\hat{\mathbf{g}}_t(\boldsymbol{\theta}) - \mathbf{g}_t(\boldsymbol{\theta})|| \leq \rho^{t-k} K_1.
\end{aligned}$$

This means that $\sup_{\boldsymbol{\theta}\in\Theta} ||\hat{\varepsilon}_t(\boldsymbol{\theta}) - \varepsilon_t(\boldsymbol{\theta})|| \xrightarrow{e.a.s} 0 \quad t \to \infty$, where by Proposition 4.3 in Krengel (1985) $\{\varepsilon_t(\boldsymbol{\theta})\}$ is a stationary and ergodic sequence. $\square$

A.4 Proof of Proposition 3

Proof: Define the norm $||\cdot||_n^{\boldsymbol{\theta}} = (\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} ||\cdot||^n)^{1/n}$, which is subadditive for $n \geq 1$. To prove that the stationary and ergodic limit sequence $\{\mathbf{g}_t\}_{t\in\mathbb{Z}}$ has $n \geq 1$ bounded moments we note that,

$$\begin{aligned}
||\mathbf{g}_t(\boldsymbol{\theta})||_n^{\boldsymbol{\theta}} &= ||\sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r C(\boldsymbol{\theta}) + \sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r B_{t+1-r}||_n^{\boldsymbol{\theta}} \\
&\leq \sum_{r=0}^{\infty} \left(\sup_{\boldsymbol{\theta}\in\Theta} ||A(\boldsymbol{\theta})^r||^n\right)^{1/n} ||C(\boldsymbol{\theta})||_n^{\boldsymbol{\theta}} + \sum_{r=0}^{\infty} \left(\sup_{\boldsymbol{\theta}\in\Theta} ||A(\boldsymbol{\theta})^r||^n\right)^{1/n} ||B_{t+1-r}||_n^{\boldsymbol{\theta}} \\
&\leq c_0 \sum_{r=0}^{\infty} \rho^r \sup_{\boldsymbol{\theta}\in\Theta} |\omega| + c_1 \sum_{r=0}^{\infty} \rho^r \left(\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} |\Delta y_{t+1-r}|^n\right)^{1/n} \\
&\leq c_2 \frac{\sup_{\boldsymbol{\theta}\in\Theta} |\omega| + \left(\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} |\Delta y_{t+1-r}|^n\right)^{1/n}}{1-\rho} < \infty.
\end{aligned}$$

By Proposition 1 we have $||\Delta y_t||_n < \infty$. The second statement of the Proposition follows from

$$||\varepsilon_t(\boldsymbol{\theta})||_n^{\boldsymbol{\theta}} = ||\Phi(\boldsymbol{\theta})\mathbf{g}_t(\boldsymbol{\theta})||_n^{\boldsymbol{\theta}} \leq \sup_{\boldsymbol{\theta}\in\Theta} ||\Phi(\boldsymbol{\theta})||_n^{\boldsymbol{\theta}} \sup_{\boldsymbol{\theta}\in\Theta} ||\mathbf{g}_t(\boldsymbol{\theta})||_n^{\boldsymbol{\theta}}. \quad \square$$

A.5 Proof of Theorem 1

Proof: The proof follows the approach of Blasques et al. (2018, Theorem 4.1) and Straumann and Mikosch (2006). In particular, we establish the following sufficient conditions:

(C1) $\hat{L}_T(\boldsymbol{\theta})$ converges almost surely to L_T uniformly over $\boldsymbol{\theta} \in \Theta$,

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|\hat{L}_T(\boldsymbol{\theta}) - L_T(\boldsymbol{\theta})\right| \xrightarrow{a.s.} 0 \text{ as } T \to \infty.$$

(C2) The limit log-likelihood contributions have one bounded moment uniformly on $\boldsymbol{\theta} \in \Theta$,

$$\mathbb{E} \sup_{\boldsymbol{\theta}\in\Theta} |l(\varepsilon(\boldsymbol{\theta}), \gamma)| < \infty.$$

(C3) $\boldsymbol{\theta}_0$ is identifiably unique,

$$\sup_{\boldsymbol{\theta}\in S^c(\boldsymbol{\theta}_0,\delta)} L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}_0),$$

where $S^c(\boldsymbol{\theta}_0, \delta)$ denotes the complement of an open ball of radius δ, centered at $\boldsymbol{\theta}_0$.

As shown in Blasques et al. (2018, Theorem 4.1), conditions C1-C3 imply that for every $\delta > 0$:

$$\lim_{T\to\infty} \sup_{\boldsymbol{\theta}\in\mathbf{B}^c(\boldsymbol{\theta}_0,\delta)} \hat{L}_T(\boldsymbol{\theta}) < L(\boldsymbol{\theta}_0)$$

and the consistency follows.

We note first that the log-likelihood takes the form,

$$\hat{L}_T(\boldsymbol{\theta}) = \frac{1}{T-k}\sum_{t=r}^{T-s} \hat{l}_t(\boldsymbol{\theta}) = \frac{1}{T-k}\sum_{t=r}^{T-s} l(\hat{\varepsilon}_t(\boldsymbol{\theta}), \boldsymbol{\gamma}) = \frac{1}{T-k}\sum_{t=r}^{T-s} \log f(\psi(L^{-1})\phi(L)\hat{g}_t(\boldsymbol{\theta}); \boldsymbol{\gamma})$$

where we have $k = r + s$ and $\hat{l}_t(\boldsymbol{\theta}) = l(\hat{\varepsilon}_t(\boldsymbol{\theta}), \boldsymbol{\gamma}) = \log f(y_t|\hat{\mu}_t(\boldsymbol{\theta}); \boldsymbol{\gamma})$ is the log-likelihood contribution of the observation at time t and $\hat{g}_t(\boldsymbol{\theta}) = y_t - \hat{\mu}_t(\boldsymbol{\theta})$ as defined before, with

$$\log f(y_t|\hat{\mu}_t(\boldsymbol{\theta}), \boldsymbol{\gamma}) = \log p_\varepsilon(\hat{\varepsilon}_t(\boldsymbol{\theta}), \boldsymbol{\gamma}) = \log\left(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(1 + \frac{\hat{\varepsilon}_t(\boldsymbol{\theta})^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right) \tag{5}$$

and filtered residuals are defined as $\hat{\varepsilon}_t(\boldsymbol{\theta}) = \psi(L^{-1})\phi(L)\hat{g}_t(\boldsymbol{\theta})$. We further let $L_T(\boldsymbol{\theta})$ denote the log-likelihood with the limit sequence $\varepsilon_t(\boldsymbol{\theta})$,

$$L_T(\boldsymbol{\theta}) = \frac{1}{T-k}\sum_{t=r}^{T-s} l_t(\boldsymbol{\theta}) = \sum_{t=r}^{T-s} l(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\gamma}),$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an iid sequence.

To prove C1, we note that by the mean value theorem

$$\begin{aligned}\hat{l}_t(\boldsymbol{\theta}) - l_t(\boldsymbol{\theta}) &= \frac{\nu+1}{2}\left[\log\left(\nu + \hat{\varepsilon}_t(\boldsymbol{\theta})^2\right) - \log\left(\nu + \varepsilon_t(\boldsymbol{\theta})^2\right)\right] \\ &= \frac{\nu+1}{2(\nu + \tilde{\varepsilon}_t(\boldsymbol{\theta})^2)}\left(\hat{\varepsilon}_t(\boldsymbol{\theta})^2 - \varepsilon_t(\boldsymbol{\theta})^2\right)\end{aligned}$$

where $\tilde{\varepsilon}$ is a point between $\hat{\varepsilon}$ and ε. Since $\tilde{\varepsilon}_t(\boldsymbol{\theta})^2$ is always positive and we assumed $\nu \geq 1$ we have $\frac{\nu+1}{2(\nu+\tilde{\varepsilon}_t(\boldsymbol{\theta})^2)} \leq 1$. Hence,

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|\hat{l}_t(\boldsymbol{\theta}) - l_t(\boldsymbol{\theta})\right| \leq \sup_{\boldsymbol{\theta}\in\Theta} \left|\hat{\varepsilon}_t(\boldsymbol{\theta})^2 - \varepsilon_t(\boldsymbol{\theta})^2\right|$$

And now since $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is SE, $\mathbb{E}\log|\varepsilon_t(\boldsymbol{\theta})| < \infty$ and $\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{\varepsilon}_t(\boldsymbol{\theta}) - \varepsilon_t(\boldsymbol{\theta})\right| \xrightarrow{e.a.s} 0$ by Lemma TA.17 of Blasques et al. (2017) we have that $\sup_{\boldsymbol{\theta}\in\Theta}\left|\hat{\varepsilon}_t(\boldsymbol{\theta})^2 - \varepsilon_t(\boldsymbol{\theta})^2\right| \xrightarrow{e.a.s} 0$. This implies,

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|\hat{l}_t(\boldsymbol{\theta}) - l_t(\boldsymbol{\theta})\right| \xrightarrow{e.a.s} 0.$$

To establish C2, we note that

$$l(\varepsilon, \boldsymbol{\theta}) = \log\left(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(1 + \frac{\varepsilon^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right),$$

and hence that,

$$\begin{aligned}
\mathbb{E}\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}|l(\varepsilon_t(\boldsymbol{\theta}),\gamma)| &= \mathbb{E}\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\left|\log\left(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(1 + \frac{\varepsilon_t(\boldsymbol{\theta})^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right)\right| \\
&\leq c_0 + \mathbb{E}\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\left|\log\left(\left(1 + \frac{\varepsilon(\boldsymbol{\theta})^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right)\right| \\
&\leq c_0 + c_1\mathbb{E}\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\left|\log\left(1 + \frac{\varepsilon_t(\boldsymbol{\theta})^2}{\nu}\right)\right| \\
&= c_0 + c_1\mathbb{E}\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\left|\varepsilon_t(\boldsymbol{\theta})\right|^{\delta} < \infty
\end{aligned}$$

for some $\delta < 1$. The last inequality follows by Proposition 3.

Condition C3 follows by noting that $L(\boldsymbol{\theta})$ exists for every $\boldsymbol{\theta} \in \Theta$, by C2. To show uniqueness of the maximizer $\boldsymbol{\theta}_0$ we need that for any $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ we have $L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}_0)$. We first show that $l\big(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0\big) = l\big(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\gamma}\big)$ almost surely if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. We know that $\varepsilon_t(\boldsymbol{\theta}_0) = \varepsilon_t$ almost surely for all t. We also know ε_t is Student's t distributed so it has a non-zero density on all $\mathbb{R}$. Hence it is enough to show that $l(h + \varepsilon; \boldsymbol{\gamma}) = l(\varepsilon; \boldsymbol{\gamma}_0)$ can hold with probability 1 if and only if $h = 0$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$. By the definition of $l(\cdot)$, for any $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$, this requires,

$$\log\left(\frac{\Gamma(\frac{\nu_1+1}{2})}{\Gamma(\frac{\nu_1}{2})\sqrt{\pi\nu_1\sigma_1^2}}\left(1 + \frac{(x+h)^2}{\sigma_1^2\nu_1}\right)^{-\frac{\nu_1+1}{2}}\right) = \log\left(\frac{\Gamma(\frac{\nu_2+1}{2})}{\Gamma(\frac{\nu_2}{2})\sqrt{\pi\nu_2\sigma_2^2}}\left(1 + \frac{x^2}{\sigma_2^2\nu_2}\right)^{-\frac{\nu_2+1}{2}}\right)$$

for all $x \in \mathbb{R}$. Clearly $l(h + \varepsilon; \boldsymbol{\gamma}) = l(\varepsilon; \boldsymbol{\gamma}_0)$ almost surely for all t requires $h = 0$ and $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2$.
We now need to prove that given that $\boldsymbol{\theta} = (\alpha, \omega, \Psi, \boldsymbol{\gamma})$ is such that $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ we can conclude that $g_t(\boldsymbol{\theta}) = g_t(\boldsymbol{\theta}_0) = v_t$ almost surely if and only if $(\alpha, \omega, \Psi) = (\alpha_0, \omega_0, \Psi_0)$. Suppose this is not the case and that $g_t(\boldsymbol{\theta}) = v_t$ almost surely for some t, than it must hold for all $t \in \mathbb{Z}$. Then we would have,

$$\begin{aligned}
g_{t+1}(\boldsymbol{\theta}) &= g_t(\boldsymbol{\theta}) - \omega - \alpha\phi(L)\psi(L^{-1})v_{t-s} + \Delta y_{t+1} \\
&= g_t(\boldsymbol{\theta}) - v_t - \big(\omega - \omega_0\big) - \alpha\sum_{h=-\infty}^{\infty}\rho_h\varepsilon_{t+h} + \alpha_0\varepsilon_{t-s} + v_{t+1} \\
&= g_t(\boldsymbol{\theta}) - v_t + \omega_0 - \omega + \alpha_0\varepsilon_{t-s} - \alpha\sum_{h=-\infty}^{\infty}\rho_h\varepsilon_{t+h} + v_{t+1}
\end{aligned}$$

Now since by hypothesis $g_t(\boldsymbol{\theta}) = v_t$ for all t then we must have:

$$\omega_0 - \omega = \alpha_0 \varepsilon_{t-s} - \alpha \sum_{h=-\infty}^{\infty} \rho_h \varepsilon_{t+h}, \quad \text{almost surely for all } t$$

Now if $\omega \neq \omega_0$ it means that the right-hand side must be a non-zero constant. But the right-hand side expression is a non degenerate function of $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ that is $\neq 0$ almost surely for all t for all $\boldsymbol{\theta} \in \Theta$ if $\alpha \neq \alpha_0$ and $\Psi \neq \Psi_0$. This means that it must be that $\omega = \omega_0$. Then since the right-hand side is non zero with probability one we can have $g_{t+1}(\boldsymbol{\theta}) = v_{t+1}$ if and only if $\alpha = \alpha_0$ and $\Psi \neq \Psi_0$.

Now that we showed that $l\big(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0\big) = l\big(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\gamma}\big)$ almost surely if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ we can use an argument on the lines of the one used in Blasques et al. (2022) using some of the arguments from Breid et al. (1991) to conclude the proof of C3. We will rely on a mean value expansion around $\boldsymbol{\theta}_0$. Recall that unfolding our limit prediction error process we have:

$$g_t(\boldsymbol{\theta}) = \sum_{i=-\infty}^{\infty} \gamma_i z_t, \tag{6}$$

with $z_t = \delta + \Delta y_t$ Moreover recall ζ_i the coefficient of the i-th element of the polynomial $\psi(L^{-1})\phi(L)$. Then consider Θ as a compact set satisfying Assumption 1-3 such that:

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta} |\zeta_i - \zeta_{i,0}| \leq C\epsilon \\
&\sup_{\boldsymbol{\theta}\in\Theta} |\gamma_i| \leq C|d|^i \\
&\sup_{\boldsymbol{\theta}\in\Theta} |\gamma_i - \gamma_{0,i}| \leq C\epsilon|d|^i \\
&\sup_{\boldsymbol{\theta}\in\Theta} |\delta - \delta_0| \leq C\epsilon
\end{aligned}$$

with $|d| < 1$. This allows us to conclude that:

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\Theta} \big|g_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta}_0)\big| &\leq \sum_{i=-\infty}^{\infty} \sup_{\boldsymbol{\theta}\in\Theta} \big|\gamma_i - \gamma_{0,i}\big| \cdot \big(\big|\Delta y_t\big| + \big|\delta_0\big|\big) + \sum_{i=-\infty}^{\infty} \sup_{\boldsymbol{\theta}\in\Theta} \big|\gamma_i\big| \cdot \sup_{\boldsymbol{\theta}\in\Theta} \big|\delta - \delta_0\big| \\
&\leq \epsilon\big(C_0 + C_1 \sum_{i=-\infty}^{\infty} |d|^i \big|z_t\big|\big) \\
\sup_{\boldsymbol{\theta}\in\Theta} \big|\varepsilon_t(\boldsymbol{\theta}) - \varepsilon_t(\boldsymbol{\theta}_0)\big| &= \sup_{\boldsymbol{\theta}\in\Theta} \big|\phi(L)\psi(L^{-1})g_t(\boldsymbol{\theta}) - \phi_0(L)\psi_0(L^{-1})g_t(\boldsymbol{\theta}_0)\big| \\
&\leq \epsilon\big(C_0 + C_1 \sum_{i=1}^{k} \big|g_{t-i}(\boldsymbol{\theta}_0)\big| + C_2 \sum_{i=-\infty}^{\infty} |d|^i \big|z_t\big|\big)
\end{aligned}$$

Moreover following Breid et al. (1991) we can write:

$$\varepsilon_t(\boldsymbol{\theta}) = \varepsilon_t(\boldsymbol{\theta}) + \varepsilon_t(\boldsymbol{\theta}_0) - \varepsilon_t(\boldsymbol{\theta}_0)$$

with:

$$\varepsilon_t(\boldsymbol{\theta}_0) - \epsilon K_t \leq \varepsilon_t(\boldsymbol{\theta}) \leq \varepsilon_t(\boldsymbol{\theta}_0) + \epsilon K_t$$

Note that the second derivatives of the log likelihood function, avoiding the repetitions in the cross derivatives, will be:

$$\frac{\partial l^2(\varepsilon(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} = \begin{cases} \frac{\partial^2 \varepsilon_t(\boldsymbol{\theta})}{\partial \phi_i \partial \phi_j} h(\varepsilon_t(\boldsymbol{\theta})) + \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \phi_i} \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \phi_j} h'(\varepsilon_t(\boldsymbol{\theta})) \\ \frac{\partial^2 \varepsilon_t(\boldsymbol{\theta})}{\partial \phi_i \partial \psi_j} h(\varepsilon_t(\boldsymbol{\theta})) + \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \phi_i} \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \psi_j} h'(\varepsilon_t(\boldsymbol{\theta})) \\ \frac{\partial^2 \varepsilon_t(\boldsymbol{\theta})}{\partial \Pi_i \partial \Pi_j} h(\varepsilon_t(\boldsymbol{\theta})) + \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \Pi_i} \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \Pi_j} h'(\varepsilon_t(\boldsymbol{\theta})) \\ \frac{\partial^2 \varepsilon_t(\boldsymbol{\theta})}{\partial \Pi_i \partial \phi_j} h(\varepsilon_t(\boldsymbol{\theta})) + \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \Pi_i} \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \phi_j} h'(\varepsilon_t(\boldsymbol{\theta})) \\ \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \Psi_i} h(\varepsilon_t(\boldsymbol{\theta})) + \sigma^{-1} \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \Psi_i} h'(\varepsilon_t(\boldsymbol{\theta})) \\ \sigma^{-1} \varepsilon_t(\boldsymbol{\theta}) h(\varepsilon_t(\boldsymbol{\theta})) + \sigma^{-2} \varepsilon_t(\boldsymbol{\theta})^2 h'(\varepsilon_t(\boldsymbol{\theta})) + 1 \end{cases}$$

Note that similarly to what has been done in the section A.6 for the proof of proposition 4 all the first and second derivatives of $\varepsilon_t(\boldsymbol{\theta})$ can be written as unfoldable and converging SREs. Unfolding these expression it is possible to show these expressions as infinite sums of the underlying z_t as in (6) with the same sequence of coefficients $\{\gamma_i\}_{t\in\mathbb{Z}}$. Then we have:

$$\sup_{\boldsymbol{\theta}\in\Theta} \left| \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} - \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_i} \right| \leq \epsilon C Z_t$$
$$\sup_{\boldsymbol{\theta}\in\Theta} \left| \frac{\partial^2 \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} - \frac{\partial^2 \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \right| \leq \epsilon C Z_t$$

where Z_t is such that $\mathbb{E}|Z_t|^n < \infty$ with n such that $\mathbb{E}|\varepsilon_t|^n < \infty$. Now that we defined bounds on these given quantities we can use the same approach as Breid et al. (1991) to conclude the proof. Here we define a mean value expansion in $\boldsymbol{\theta}_0$ of our expected likelihood difference.

$$\begin{aligned} &\mathbb{E}\big[l(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\theta}) - l(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0)\big] \\ &= \mathbb{E}\Bigg[\sum_{i=1}^{k} \frac{\partial l(\varepsilon(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_i} \big(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i,0}\big) + \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\partial l^2(\varepsilon(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \big(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i,0}\big)\big(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j,0}\big) \\ &+ \sum_{i=1}^{k}\sum_{j=1}^{k} \left(\frac{\partial l^2(\varepsilon(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} - \frac{\partial l^2(\varepsilon(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \right) \big(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i,0}\big)\big(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j,0}\big) \Bigg] \end{aligned} \tag{7}$$

From now on we will provide an argument for the derivative taken with respect to $i \leq r$ but the same argument holds for the non-causal part. Note that using $\varepsilon_t(\boldsymbol{\theta}_0) = \varepsilon_t$ we have:

$$\mathbb{E}\left[\frac{\partial l(\varepsilon(\boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta}_i} \big(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i,0}\big) \right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\partial l(\varepsilon(\boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta}_i} \big(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i,0}\big) \middle| \mathcal{F}_{t-1} \right] \right] = 0$$

For what concerns the third term we have:

$$\mathbb{E}\Bigg[\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\partial l^2(\varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \big(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i,0}\big)\big(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j,0}\big) \Bigg] = -\big(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i,0}\big)' \mathcal{I}(\boldsymbol{\theta}_0) \big(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i,0}\big)$$

Finally for the last term we can apply a similar reasoning as what it is done in Breid et al. (1991). We have:

$$
\begin{aligned}
\mathbb{E}\sup_{\boldsymbol{\theta}}\Bigg|&\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i}\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_j}h'(\varepsilon_t(\boldsymbol{\theta}^*))+\frac{\partial^2\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i\partial\phi_j}h(\varepsilon_t(\boldsymbol{\theta}^*))\\
&-\frac{\partial\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_i}\frac{\partial\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_j}h'(\varepsilon_t(\boldsymbol{\theta}_0))-\frac{\partial^2\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_i\partial\phi_j}h(\varepsilon_t(\boldsymbol{\theta}_0))\Bigg|\\
\leq&\mathbb{E}\sup_{\boldsymbol{\theta}}\left|\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i}\frac{\partial\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_j}h'(\varepsilon_t(\boldsymbol{\theta}_0))-\frac{\partial\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_i}\frac{\partial\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_j}h'(\varepsilon_t(\boldsymbol{\theta}_0))\right|\\
&+\mathbb{E}\sup_{\boldsymbol{\theta}}\left|\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i}\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_j}h'(\varepsilon_t(\boldsymbol{\theta}_0))-\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i}\frac{\partial\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_j}h'(\varepsilon_t(\boldsymbol{\theta}_0))\right|\\
&+\mathbb{E}\sup_{\boldsymbol{\theta}}\left|\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i}\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_j}h'(\varepsilon_t(\boldsymbol{\theta}^*))-\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i}\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_j}h'(\varepsilon_t(\boldsymbol{\theta}_0))\right|\\
&+\mathbb{E}\sup_{\boldsymbol{\theta}}\left|\frac{\partial^2\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i\partial\phi_j}h(\varepsilon_t(\boldsymbol{\theta}_0))-\frac{\partial^2\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_i\partial\phi_j}h(\varepsilon_t(\boldsymbol{\theta}_0))\right|\\
&+\mathbb{E}\sup_{\boldsymbol{\theta}}\left|\frac{\partial^2\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i\partial\phi_j}h(\varepsilon_t(\boldsymbol{\theta}^*))-\frac{\partial^2\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i\partial\phi_j}h(\varepsilon_t(\boldsymbol{\theta}_0))\right|\\
&=c_1+c_2+c_3+c_4+c_5
\end{aligned}
$$

Then:

$$
c_1\leq\epsilon C\mathbb{E}\left|Z_t\frac{\partial\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_j}h'(\varepsilon_t(\boldsymbol{\theta}_0))\right|\to 0,\quad\text{as }\epsilon\to 0
$$

Note that $\mathbb{E}\left|Z_t\frac{\partial\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_j}h'(\varepsilon_t(\boldsymbol{\theta}_0))\right|<\infty$ as $\frac{\partial\varepsilon_t(\boldsymbol{\theta}_0)}{\partial\phi_j}$and $h'(\varepsilon_t(\boldsymbol{\theta}_0))$ are independent and it is possible to split the infinite past and future elements in Z_t such that all the elements in the expectation are bounded by $\mathbb{E}|\varepsilon_t|^2<\infty$.
By a similar argument also $c_2\to 0$ as $\epsilon\to 0$. Moreover as in Breid et al. (1991) we can split:

$$
h'(x)=h_1(x)-h_2(x)
$$

with $h_i(\cdot)$ non-decreasing functions such that:

$$
h_i(x)=O(|x|^k),\quad\text{as }|x|\to\infty
$$

with k such that $\mathbb{E}|\varepsilon_t|^{2+k}<\infty$. Note also that the same operation is possible for $h(x)$. With this definition we can define:

$$
X_{i,t}=\begin{cases}
h_i\left(\frac{\varepsilon_t(\boldsymbol{\theta}_0)-\varepsilon CK_t}{\sigma_0-\epsilon}\right)-h_i\left(\frac{\varepsilon_t(\boldsymbol{\theta}_0)+\varepsilon CK_t}{\sigma_0+\epsilon}\right), & \text{if } \varepsilon_t(\boldsymbol{\theta}_0)+\varepsilon CK_t¿0\\
h_i\left(\frac{\varepsilon_t(\boldsymbol{\theta}_0)-\varepsilon CK_t}{\sigma_0+\epsilon}\right)-h_i\left(\frac{\varepsilon_t(\boldsymbol{\theta}_0)+\varepsilon CK_t}{\sigma_0-\epsilon}\right), & \text{if } \varepsilon_t(\boldsymbol{\theta}_0)+\varepsilon CK_t¡0\\
h_i\left(\frac{\varepsilon_t(\boldsymbol{\theta}_0)-\varepsilon CK_t}{\sigma_0-\epsilon}\right)-h_i\left(\frac{\varepsilon_t(\boldsymbol{\theta}_0)+\varepsilon CK_t}{\sigma_0-\epsilon}\right), & \text{otherwise}
\end{cases}
$$

Then we can bound:

$$
\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}c_3\leq\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_i}\frac{\partial\varepsilon_t(\boldsymbol{\theta}^*)}{\partial\phi_j}\Big(X_{1,t}+X_{2,t}\Big)
$$

Using the moment bounds it is possible to show that this expected value is finite, then by dominated convergence we have that $c_3\to 0$ as $\varepsilon\to 0$. We can apply the same approach to c_4 and

c_5 so that we showed that the difference between the second derivatives in the last term of (7) goes to zero with ϵ for $i, j \leq k$. The reasoning is similar for other elements of the second derivative of the score as argued in Breid et al. (1991), hence we have that:

$$\mathbb{E}\Big[\sum_{i=1}^{k}\sum_{j=1}^{k}\left(\frac{\partial l^2(\varepsilon(\boldsymbol{\theta}^*))}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}-\frac{\partial l^2(\varepsilon(\boldsymbol{\theta}_0))}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\right)\big(\boldsymbol{\theta}_i-\boldsymbol{\theta}_{i,0}\big)\big(\boldsymbol{\theta}_j-\boldsymbol{\theta}_{j,0}\big)\Big]\to 0,\quad \text{for } \epsilon\to 0$$

so that there is a $\varepsilon > 0$ such that for all $\boldsymbol{\theta} \in \Theta$ such that $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ we have:

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_0) = \mathbb{E}\big[l(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\gamma}) - l(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)\big] < 0$$

Moreover assumption 4 ensures that $\boldsymbol{\theta}$ is a compact set and the uniform convergence result showed in part **i)** implies the continuity of the limit criterion function $L(\boldsymbol{\theta})$. These two results combined with the uniqueness of the maximizer imply the result. □

A.6 Proof of Proposition 4

Proof: We establish (a) by noting that the first derivative of the limit process for our prediction errors, defined in (13), takes the form

$$\frac{\partial \mathbf{g}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{r=0}^{\infty}\frac{\partial C_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}A(\boldsymbol{\theta})^r + \sum_{r=0}^{\infty}K_r(\boldsymbol{\theta})C(\boldsymbol{\theta}) + \sum_{r=0}^{\infty}K_r(\boldsymbol{\theta})B_{t+1-r} \tag{8}$$

where $K_r(\boldsymbol{\theta}) = \sum_{k=1}^{r}\mathbf{A}^{k-1}(\boldsymbol{\theta})\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{A}(\boldsymbol{\theta})^{r-k}$. Further, we note that,

$$\sup_{\boldsymbol{\theta}\in\Theta}||K_r(\boldsymbol{\theta})|| \leq \sum_{k=1}^{r}||\mathbf{A}^{k-1}(\boldsymbol{\theta})\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{A}(\boldsymbol{\theta})^{r-k}||^{\boldsymbol{\theta}} \leq rK\rho^{r-1}\big|\big|\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\big|\big|^{\boldsymbol{\theta}}.$$

Taking the derivative the unfolded filtered prediction error, defined in (4), we get instead,

$$\frac{\partial\hat{\mathbf{g}}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{r=0}^{t-k}\mathbf{A}(\boldsymbol{\theta})^r\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} + \sum_{r=0}^{t-k}K_r(\boldsymbol{\theta})C(\boldsymbol{\theta}) + \sum_{r=0}^{t-k}K_r(\boldsymbol{\theta})B_{t+1-r} + K_{t-k}(\boldsymbol{\theta})\hat{\mathbf{g}}_k. \tag{9}$$

Hence, the difference is given by,

$$\begin{aligned}
&\Big|\Big|\frac{\partial\hat{\mathbf{g}}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \frac{\partial\mathbf{g}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|\Big|^{\boldsymbol{\theta}} \\
&= \Big|\Big|\sum_{r=t-k+1}^{\infty}\left(\mathbf{A}(\boldsymbol{\theta})^r\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} + K_r(\boldsymbol{\theta})C(\boldsymbol{\theta}) + K_r(\boldsymbol{\theta})B_{t+1-r}\right) + K_{t-k}(\boldsymbol{\theta})\hat{\mathbf{g}}_k\Big|\Big|^{\boldsymbol{\theta}} \\
&\leq \sum_{r=t-k+1}^{\infty}\Big|\Big|\left(\mathbf{A}(\boldsymbol{\theta})^r\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\Big|\Big|^{\boldsymbol{\theta}} + \sum_{r=t-k+1}^{\infty}\big|\big|K_r(\boldsymbol{\theta})\big|\big|^{\boldsymbol{\theta}}\cdot\Big|\Big|C(\boldsymbol{\theta}) + B_{t+1-r}\Big|\Big|^{\boldsymbol{\theta}} + \big|\big|K_{t-k}(\boldsymbol{\theta})\big|\big|^{\boldsymbol{\theta}}\cdot\big|\big|\hat{\mathbf{g}}_k\big|\big| \\
&\leq \sum_{r=t-k+1}^{\infty}\rho^r K_1 + \sum_{r=t-k+1}^{\infty}r\rho^r K_2\cdot\left(|\omega|^{\boldsymbol{\theta}} + \big|\big|B_{t+1-r}\big|\big|^{\boldsymbol{\theta}}\right) + (t-k)\rho^{t-k-1}K_3\big|\big|\hat{\mathbf{g}}_k\big|\big| \\
&\leq (t-k)\rho^{t-k}\left[\frac{K_1}{t-k}\frac{\rho}{1-\rho} + K_2\left(\frac{\rho|\omega|^{\boldsymbol{\theta}}}{(1-\rho)^2} + \sum_{j=1}^{\infty}j\rho^j\big|\big|B_{k-j}\big|\big|^{\boldsymbol{\theta}}\right) + \rho^{-1}\big|\big|\hat{\mathbf{g}}_k\big|\big|^{\boldsymbol{\theta}}\right]
\end{aligned}$$

where the inequality follows from the fact that we know under our assumptions that $\sup_{\boldsymbol{\theta}\in\Theta}\big|\big|A(\boldsymbol{\theta})^r\big|\big| \leq K\rho^r$ and $c = \sup_{\boldsymbol{\theta}\in\Theta}||\partial\mathbf{A}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|| < \infty$ because it is a continuous function over a compact set. Naturally, this means that,

$$\big|\big|\frac{\partial\hat{\mathbf{g}}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \frac{\partial\mathbf{g}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\big|\big| \xrightarrow{e.a.s.} 0 \quad \text{as} \quad t\to\infty$$

Finally, taking the expression in (8) we have, by subadditivity of $||\cdot||^n_{\boldsymbol{\theta}}$, that,

$$\begin{aligned}\mathbb{E}\sup_{\boldsymbol{\theta}}\big|\big|\frac{\partial\mathbf{g}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\big|\big|^n &= \mathbb{E}\sup_{\boldsymbol{\theta}}\big|\big|\sum_{r=0}^{\infty}\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}A(\boldsymbol{\theta})^r + \sum_{r=0}^{\infty}K_r(\boldsymbol{\theta})C(\boldsymbol{\theta}) + \sum_{r=0}^{\infty}K_r(\boldsymbol{\theta})B_{t+1-r}\big|\big|^n \\ &\leq b_1 + b_2\big(|\omega|^n_{\boldsymbol{\theta}} + \mathbb{E}|\Delta y_{t+1-r}|^n\big) < \infty\end{aligned}$$

under the assumptions that $\mathbb{E}|\varepsilon_t|^n < \infty$, with $b_1 = \frac{K_1}{1-\rho^n}$ and $b_2 = \frac{K_2\rho^n}{(1-\rho^n)^2}$.

We now show that (b) holds by taking the derivative of the expression in (8) to get the limit second derivative of the limit process of the prediction erros, that takes the form

$$\frac{\partial^2\mathbf{g}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \sum_{r=0}^{\infty}\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}K_r(\boldsymbol{\theta}) + \sum_{r=0}^{\infty}\left(Q_r(\boldsymbol{\theta})C(\boldsymbol{\theta}) + K_r(\boldsymbol{\theta})\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right) + \sum_{r=0}^{\infty}Q_r(\boldsymbol{\theta})B_{t+1-r} \tag{10}$$

where we have

$$\begin{aligned}Q_r(\boldsymbol{\theta}) = \sum_{k=1}^{r}\Bigg(\sum_{j=1}^{k-1}\mathbf{A}^{j-1}(\boldsymbol{\theta})\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{A}(\boldsymbol{\theta})^{k-1-j}\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{A}(\boldsymbol{\theta})^{r-k} + \mathbf{A}(\boldsymbol{\theta})^{k-1}\frac{\partial^2\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathbf{A}(\boldsymbol{\theta})^{r-k} + \\ \sum_{j=1}^{r-k}\mathbf{A}^{k-1}(\boldsymbol{\theta})\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{A}(\boldsymbol{\theta})^{j-1}\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{A}(\boldsymbol{\theta})^{r-k-j}\Bigg).\end{aligned}$$

Now, we note that

$$\begin{aligned}||Q_r(\boldsymbol{\theta})||^{\boldsymbol{\theta}} &\leq \sum_{k=1}^{r}\left((r-1)K\rho^{r-2}\big|\big|\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\big|\big|^2_{\boldsymbol{\theta}} + K\rho^{r-1}\big|\big|\frac{\partial^2\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\big|\big|^{\boldsymbol{\theta}}\right) \\ &\leq r^2K\rho^{r-2}\left(\big|\big|\frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\big|\big|^2_{\boldsymbol{\theta}} + \rho\big|\big|\frac{\partial^2\mathbf{A}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\big|\big|^{\boldsymbol{\theta}}\right).\end{aligned}$$

While by taking the derivative of the unfolded filtered derivative, defined in (9), we have,

$$\frac{\partial^2\hat{\mathbf{g}}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \sum_{r=0}^{t-k}K_r(\boldsymbol{\theta})\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} + \sum_{r=0}^{t-k}\left(Q_r(\boldsymbol{\theta})C(\boldsymbol{\theta}) + K_r(\boldsymbol{\theta})\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right) + \sum_{r=0}^{t-k}Q_r(\boldsymbol{\theta})B_{t+1-r} + Q_{t-k}(\boldsymbol{\theta})\hat{\mathbf{g}}_k \tag{11}$$

Then by similar arguments as in point (a) we have,

$$\begin{aligned}\left|\left|\frac{\partial^2\hat{\mathbf{g}}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial^2\mathbf{g}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|\right|^{\boldsymbol{\theta}} &= \left|\left|\sum_{r=t-k+1}^{\infty}\left(2K_r(\boldsymbol{\theta})^r\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} + Q_r(\boldsymbol{\theta})\big[C(\boldsymbol{\theta}) + B_{t+1-r}\big]\right) + Q_{t-k}(\boldsymbol{\theta})\hat{\mathbf{g}}_k\right|\right|^{\boldsymbol{\theta}} \\ &\leq K_0(t-k+1)\rho^{t-k} + (t-k+1)^2\rho^{t-k-1}K_1 + (t-k)^2\rho^{t-k-2}K_2||\hat{\mathbf{g}}_k||^{\boldsymbol{\theta}} \\ &\leq (t-k)^2\rho^{t-k}\left(K_{0,t} + \rho^{-1}K_{1,t} + \rho^{-2}K_2\big|\big|\hat{\mathbf{g}}_k\big|\big|\right).\end{aligned}$$

As a result,

$$||\frac{\partial^2 \hat{\mathbf{g}}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial^2 \mathbf{g}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}|| \xrightarrow{e.a.s.} 0 \quad \text{as} \quad t \to \infty$$

Hence, we can use (10) and in a similar way to (a), by the sub-additivity of the norm and by the fact that $\mathbb{E}|\varepsilon_t|^n < \infty$ we have,

$$\begin{aligned}
&\mathbb{E}\sup_{\boldsymbol{\theta}}||\frac{\partial^2 \mathbf{g}_{t+1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}||^n \\
&\quad = \mathbb{E}\sup_{\boldsymbol{\theta}}||\sum_{r=0}^{\infty}\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}K_r(\boldsymbol{\theta}) + \sum_{r=0}^{\infty}\left(Q_r(\boldsymbol{\theta})C(\boldsymbol{\theta}) + K_r(\boldsymbol{\theta})\frac{\partial C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right) + \sum_{r=0}^{\infty}Q_r(\boldsymbol{\theta})B_{t+1-r}||^n \\
&\quad \leq c_1 + c_2\big(K_2 + K_3\mathbb{E}\big|B_{t+1-r}\big|^n\big) < \infty
\end{aligned}$$

with $c_1 = \frac{\rho^n K_3}{(1-\rho^n)^2}$ and $c_2 = \frac{\rho^n(1+\rho^n)}{(1-\rho^n)^3}$. $\square$

A.7 Proof of Theorem 2

Proof: We follow the argument of Theorem 3.1 of Gorgi and Koopman (2021) and Section 7 of Straumann and Mikosch (2006). We first start showing the normality of the ML estimator which relies exclusively on the limit likelihood function, defined by,

$$\tilde{\boldsymbol{\theta}}_T = \arg\max_{\boldsymbol{\theta}\in\Theta} L_T(\boldsymbol{\theta}) \tag{12}$$

The final result is then proved by showing that $\sqrt{T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ and $\sqrt{T}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ have the same asymptotic distribution.

Note that $L_T(\boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$, the expression for L'_T and L''_T are available in the Appendix [6]. By Proposition 4.3 in Krengel (1985), we have that L'_T and L''_T are both stationary and ergodic since they are continuous function of $g_t(\boldsymbol{\theta})$, $\partial g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\partial^2 g_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ that are stationary and ergodic by Proposition 2. Joint stationarity and ergodicity of the three sequences can be recovered by the stable dependence on the underlying sequence $\{\varepsilon_t\}_{t\in\mathbb{Z}}$. Now, application of the mean value theorem yields,

$$L_T^{'}(\tilde{\boldsymbol{\theta}}_T) = L_T^{'}(\boldsymbol{\theta}_0) + L_T^{''}(\boldsymbol{\theta}^*)(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}^*$ is a point between $\boldsymbol{\theta}_0$ and $\tilde{\boldsymbol{\theta}}_T$. By the definition given in (12) we have that $L_T^{'}(\tilde{\boldsymbol{\theta}}_T) = 0$, this means that we have,

$$L_T^{''}(\boldsymbol{\theta}_T^*)\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) = -\sqrt{T}L_T^{'}(\boldsymbol{\theta}_0).$$

Now we have that $\tilde{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$, with $\boldsymbol{\theta}_0$ in the interior of $\boldsymbol{\theta}$ by assumption. We also have that under Assumptions 1-4 and if assumption 3 holds with $n > 4$ we can apply lemma 1 such that $L_T^{''}(\boldsymbol{\theta}_T^*)$ has a uniformly bounded moment. Then by the ergodic theorem of Rao (1962) we have that,

$$-L_T^{''}(\boldsymbol{\theta}_T^*) \xrightarrow{a.s.} -\mathbb{E}[l_t''(\boldsymbol{\theta}_0)].$$

We also have, by Lemma 2, that $\mathbb{E}||l_t''(\boldsymbol{\theta}_0)||$ is positive definite, and by Lemma 3 we have $\sqrt{T}L'_T(\boldsymbol{\theta}_0) \xrightarrow{d} N(0,\Omega^{-1})$ as $T\to\infty$. Therefore we conclude that

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0,\Omega) \quad \text{as} \quad T\to\infty$$

[6] see Appendix E

where $\Omega = -\mathbb{E}[l_t''(\boldsymbol{\theta}_0)]^{-1}$. Finally, we show that $\hat{\boldsymbol{\theta}}_T$ and $\tilde{\boldsymbol{\theta}}_T$ have the same asymptotic distribution. We can do this as in Straumann and Mikosch (2006) by noting that,

$$L_T^{'}(\hat{\boldsymbol{\theta}}_T) = L_T^{'}(\tilde{\boldsymbol{\theta}}_T) + L_T^{''}(\boldsymbol{\theta}^{**})(\tilde{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T)$$

where through abuse of notation, we let $\boldsymbol{\theta}^{**}$ denote a point between $\tilde{\boldsymbol{\theta}}_T$ and $\hat{\boldsymbol{\theta}}_T$, rowise in the score vector. Now we have that by definition $L_T^{'}(\tilde{\boldsymbol{\theta}}_T) = 0$ and $\hat{L}_T^{'}(\hat{\boldsymbol{\theta}}_T) = 0$. Then the previous equation is equivalent to,

$$\sqrt{T}\big(\hat{L}_T^{'}(\hat{\boldsymbol{\theta}}_T) - L_T^{'}(\hat{\boldsymbol{\theta}}_T)\big) = L_T^{''}(\boldsymbol{\theta}^{**})\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T)$$

The left hand term goes to zero a.s. by Lemma 4 as $T \to \infty$. We also have that $\boldsymbol{\theta}_T^{**} \xrightarrow{a.s.} \boldsymbol{\theta}_0$ and as before by Lemma 1 we have that $L_T^{''}(\boldsymbol{\theta}^{**}) \xrightarrow{a.s.} \mathbb{E}[l_t''(\boldsymbol{\theta}_0)]$ as $T \to \infty$. This implies that $\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T) \xrightarrow{a.s.} 0$ implying that $\tilde{\boldsymbol{\theta}}_T$ and $\hat{\boldsymbol{\theta}}_T$ have the same distribution. $\square$

B Lemmas

Lemma 1 *Under assumption 1-4 and if assumption 3 holds with $n > 4$ we have that the second derivative of the loglikelihood function has a uniformly bounded moment, that is $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}||l_t''(\boldsymbol{\theta})|| < \infty$.*

Proof: We have that $l_t''(\boldsymbol{\theta}) = \partial l_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$. First recall the division of the parameter vector $\boldsymbol{\theta} = (\alpha, \omega, \boldsymbol{\gamma})$, where α and ω are the parameters driving only the update equation. Using the definition we have in the Appendix (18) this means,

$$\begin{aligned}\mathbb{E}||l_t''(\boldsymbol{\theta})||^{\boldsymbol{\theta}} \leq \mathbb{E}\left|\left|\frac{\partial^2 l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|\right|^{\boldsymbol{\theta}} &+ \mathbb{E}\left|\left|\frac{\partial^2 l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial\boldsymbol{\theta}\partial\varepsilon}\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|^{\boldsymbol{\theta}} \\ &+ \mathbb{E}\left|\left|\frac{\partial^2 l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial\varepsilon\partial\varepsilon}\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\varepsilon_t(\boldsymbol{\theta})'}{\partial\boldsymbol{\theta}}\right|\right|^{\boldsymbol{\theta}} \\ &+ \mathbb{E}\left|\left|\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial^2 l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial\varepsilon\partial\boldsymbol{\theta}}\right|\right|^{\boldsymbol{\theta}} + \mathbb{E}\left|\left|\frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial\varepsilon}\frac{\partial^2\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|\right|^{\boldsymbol{\theta}}.\end{aligned}$$

The first term is defined by (19), and we can see in (22) that it is bounded. To show that the next terms are bounded we use the generalized Holder inequality saying that for the norm $||\cdot||_p = (\mathbb{E}||\cdot||^p)^{1/p}$ for random variables or random vectors x and y we have $||x\cdot y|| \leq ||x||_p||y||_q$ with $p, q > 0$ such that $pq/(p+q) = 1$. For the second terms, and symmetrically for fourth term, this implies,

$$\mathbb{E}\left|\left|\frac{\partial^2 l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial\boldsymbol{\theta}\partial\varepsilon}\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|^{\boldsymbol{\theta}} \leq \mathbb{E}\left|\left|\frac{\partial^2 l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial\boldsymbol{\theta}\partial\varepsilon}\right|\right|_2^{\boldsymbol{\theta}}\mathbb{E}\left|\left|\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|_2^{\boldsymbol{\theta}} < \infty,$$

where again the first term is defined in (19) in the appendix and bounded by the expression in (22). Moreover under the assumption of a finite moment for $n = 4$ the second term is bounded by Lemma 5.

The same approach holds for the third and the last term,

$$\begin{aligned}\mathbb{E}\left|\left|\frac{\partial^2 l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial\varepsilon\partial\varepsilon}\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|^{\boldsymbol{\theta}} &\leq \mathbb{E}\left|\left|\frac{\partial^2 l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial\varepsilon\partial\varepsilon}\right|\right|_2^{\boldsymbol{\theta}}\mathbb{E}\left|\left|\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|_4^{\boldsymbol{\theta}}\mathbb{E}\left|\left|\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|_4^{\boldsymbol{\theta}} < \infty \\ \mathbb{E}\left|\left|\frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial\varepsilon}\frac{\partial^2\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|\right|^{\boldsymbol{\theta}} &< \mathbb{E}\left|\left|\frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}),\gamma)}{\partial\varepsilon}\right|\right|_2^{\boldsymbol{\theta}}\mathbb{E}\left|\left|\frac{\partial^2\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|\right|_2^{\boldsymbol{\theta}} < \infty\end{aligned}$$

Lemma 2 $\mathbb{E}||l_t^{''}(\boldsymbol{\theta}_0)||$ *is positive definite.*

Proof First recall the division of the parameter vector $\boldsymbol{\theta} = (\alpha, \omega, \Psi, \boldsymbol{\gamma})$, where α and ω are the parameters driving only the update equation. For the purpose of this proof let us include the MAR parameter in the distribution parameters vector, so $\boldsymbol{\gamma} = (\phi_1, ..., \phi_r, \psi_1, ..., \psi_s, \sigma, \nu)$, as we are interested in the difference between the filter parameters and the other ones. We note that under Assumption 2 we have $-\mathbb{E}\big[l_t''(\boldsymbol{\theta}_0)\big] = \mathbb{E}\big[l_t'(\boldsymbol{\theta}_0)l_t'(\boldsymbol{\theta}_0)'\big]$ by Fischer information matrix equality. We are assuming that the model is correctly specified such that $l_t(\boldsymbol{\theta}_0)$ is the true log density evaluated at y_t. We also note that $l_t(\boldsymbol{\theta}_0)$ is twice continuosly differentiable and that the second derivative of $l_t(\boldsymbol{\theta}_0)$ has a bounded moment according to Lemma 1. Then the equality follows using standard arguments.

Now to show that the matrix is invertible we note that $\mathbb{E}\big[l_t'(\boldsymbol{\theta}_0)l_t'(\boldsymbol{\theta}_0)'\big]$ is positive semi-definite by construction so what is left to prove is that it is also not singular. We then have to prove that,

$$v'l_t'(\boldsymbol{\theta}) = 0 \quad \text{a.s if and only if } v = 0$$

where $v \in \mathbb{R}^{k+4}$. We can express our first derivative as:

$$v' l_t'(\boldsymbol{\theta}) = v' \left[\begin{pmatrix} 0 \\ 0 \\ \frac{\partial l_t'(\varepsilon(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} \end{pmatrix} + \begin{pmatrix} \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \alpha} \\ \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \omega} \\ \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma}} \end{pmatrix} \frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\gamma}_0)}{\partial \varepsilon} \right].$$

We can then split the vector $v = (v_1, v_2)$ with $v_1 \in \mathbb{R}^2$, that is the vector of elements corresponding to the zeros in the first term of the derivative of the likelihood, and with $v_2 \in \mathbb{R}^{k+2}$ that is the rest of the elements.

It is possible to argue as in the proof of Gorgi and Koopman (2021) that we could have $v' l_t'(\boldsymbol{\theta}) = 0$ almost surely with $v \neq 0$ only with the following cases: (i) $v_1 \neq 0$ and $v_2 = 0$, (ii) $v_1 = 0$ and $v_2 \neq 0$ and (iii) $v_1 \neq 0$ and $v_2 \neq 0$. For the first case (i) to hold we must have,

$$\frac{\partial l_t(\varepsilon_t, \boldsymbol{\gamma}_0)}{\partial \varepsilon} v_1' \begin{pmatrix} \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \alpha} \\ \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \omega} \end{pmatrix} = 0.$$

Observing the score function in the Appendix E we can see that $\partial l_t(\varepsilon_t, \boldsymbol{\gamma}_0)/\partial \varepsilon \neq 0$ with probability 1. Then for this to hold we need to have $v_{1,1} \partial \varepsilon_t(\boldsymbol{\theta}_0)/\partial \alpha + v_{1,2} \partial \varepsilon_t(\boldsymbol{\theta}_0)/\partial \omega = 0$ a.s. We have that,

$$\mathbf{g}_{t+1}(\boldsymbol{\theta}) = \sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r C(\boldsymbol{\theta}) + \sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r B_{t+1-r} \tag{13}$$

$$\frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \omega} = \Phi(\boldsymbol{\theta}) \frac{\partial \mathbf{g}_t(\boldsymbol{\theta})}{\partial \omega} = \Phi(\boldsymbol{\theta}) \Big(\sum_{r=0}^{\infty} A(\boldsymbol{\theta})^r \frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big)$$

$$\frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \alpha} = \Phi(\boldsymbol{\theta}) \frac{\partial \mathbf{g}_t(\boldsymbol{\theta})}{\partial \alpha} = \Phi(\boldsymbol{\theta}) \sum_{r=0}^{\infty} \Big(\sum_{k=1}^{r-1} A(\boldsymbol{\theta})^k \frac{\partial A(\boldsymbol{\theta})}{\partial \alpha} A(\boldsymbol{\theta})^{r-k} \big[C(\boldsymbol{\theta}) + B_{t+1-r} \big] \Big)$$

Now noting that all the elements, including the derivative of $A(\boldsymbol{\theta})$ are well defined, we know that $C(\boldsymbol{\theta}) + B_{t+1-r} = \frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ if and only if $\Delta y_{t+1-r} - \omega = -1$ and this happens with probability zero. Since these two terms are different with probability 1 we have that the derivative processes are linearly independent. This means option (i) can not hold.
Now we need to rule out case (ii). In this case:

$$v_2' \Big[\frac{\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} \frac{\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \varepsilon}^{-1} + \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma}} \Big] = 0$$

We first note that $\partial \varepsilon_t(\boldsymbol{\theta}_0)/\partial \boldsymbol{\gamma}_i = 0$ for $i = k+2$ and on the other hand $\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)/\partial \boldsymbol{\gamma}_i = 0$ for $i \neq k+1, k+2$. Define $j = \{k+1, k+2\}$. We can then repeat our splitting argument, one would need to find a $\mathbf{v}_{2,1}, \mathbf{v}_{2,2}$ such that,

$$\mathbf{v}_{2,1} \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma}_{-j}} + \mathbf{v}_{2,2} \frac{\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_j} \frac{\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \varepsilon}^{-1} = 0 \quad \text{a.s.}$$

We would need either $v_{2,1} \neq 0$ and $v_{2,2} = 0$, (ii) $v_{2,1} = 0$ and $v_{2,2} \neq 0$ and (iii) $v_{2,1} \neq 0$ and $v_{2,2} \neq 0$. Let us start with the first expression.

At the same time we can have a look at the derivatives of the first term, we can write $\Phi(\boldsymbol{\theta}) = (\varphi_0, ..., \varphi_{k-1})$ where φ_i is the coefficient of z^{i-s} in $\frac{1}{\sigma}(1 - \phi_1 z - \phi_2 z^2 - ... - \phi_r z^r)(1 - \psi_1 z^{-1} - \psi_2 z^{-2} -$

$... - \psi_s z^{-s})$. Then:

$$\varphi_i = \sum_{j=1}^{i} \psi_{s-j+1}\phi_{i-j} \quad \text{for } i = 1, ..., s$$
$$\varphi_i = \sum_{j=1}^{k-i+1} \psi_{j-1}\phi_{j+i-s-1} \quad \text{for } i = s+1, ..., k.$$

Then,

$$\frac{\partial \Phi(\boldsymbol{\theta})}{\partial \phi_1} = \frac{1}{\sigma}\big[0, \psi_s ..., \psi_1, 1, 0, ..., 0\big]$$
$$\frac{\partial \Phi(\boldsymbol{\theta})}{\partial \phi_r} = \frac{1}{\sigma}\big[0, 0, ..., 0, \psi_s ..., \psi_1, 1\big]$$
$$\frac{\partial \Phi(\boldsymbol{\theta})}{\partial \psi_s} = \frac{1}{\sigma}\big[1, \phi_1, ..., \phi_r, 0, 0, ..., 0\big]$$
$$\frac{\partial \Phi(\boldsymbol{\theta})}{\partial \psi_1} = \frac{1}{\sigma}\big[0, .., 0, 1, \phi_1, ..., \phi_r, , 0\big]$$

at the same time we have that $\frac{\partial A(\boldsymbol{\theta})}{\partial \gamma_i}$ is a matrix with $\alpha \frac{\partial \Phi(\boldsymbol{\theta})}{\partial \gamma_i}$ on the first row and zeros everywhere else. This means that,

$$\frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma}_i} = \frac{\partial \Phi(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_i} g_t(\boldsymbol{\theta}) + \Phi(\boldsymbol{\theta}) \sum_{r=1}^{\infty} \left(\sum_{k=1}^{r-1} A(\boldsymbol{\theta})^k \frac{\partial A(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_i} A(\boldsymbol{\theta})^{r-k} \big[C(\boldsymbol{\theta}) + B_{t+1-r}\big] \right)$$

so every element of this vector is a linear function of linearly independent vectors.

This means that it is not possible to find a non-zero vector such that,

$$\mathbf{v}_{2,1} \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma}_{-j}} = 0 \quad \text{a.s.}$$

Now we can move to the second case, in the appendix we can find an expression for the last term. Both its elements are non degenerate so it is non-zero with positive probability. This means we would need $\mathbf{v}_{2,2} = 0$ to have,

$$\mathbf{v}_{2,2} \frac{\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_j} \frac{\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \varepsilon}^{-1} = 0 \quad \text{a.s.}$$

For the last case we would need to find $\mathbf{v}_{2,1}, \mathbf{v}_{2,2} \neq 0$ such that,

$$\mathbf{v}_{2,1} \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma}_{-j}} = -\mathbf{v}_{2,2} \frac{\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_j} \frac{\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \varepsilon}^{-1} \quad \text{a.s.}$$

But we have that our right term depends on ε_t while the one on the left is $\mathcal{F}_{t-1}$-measurable, with all the elements being not degenerate, so this equation cannot hold. This rules out the possibility that $v_2 = 0$ and $v_1 \neq 0$.

Then for the third case (iii) we would need,

$$v_1' \begin{pmatrix} \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \alpha} \\ \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \omega} \end{pmatrix} + v_2' \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma}} = v_2' \frac{\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}} \frac{\partial l_t'(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma}_0)}{\partial \varepsilon}^{-1} \quad \text{a.s.}$$

Now the left hand side is $\mathcal{F}_{t-1}$-measurable and the right hand side is not because it depends on ε_t, and since we have that all the derivatives are non degenerate we have that this equation cannot hold a.s. By this we can conclude our proof.

Lemma 3 *Under the assumptions 1- 4- 3 and if assumption 2 holds with $n \geq 4$ we have that,*

$$\sqrt{T} L'_T(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, K) \quad \textit{with } T \to \infty$$

where $L'_T(\boldsymbol{\theta}_0) = \sum_{t=1}^{T} l'_t(\boldsymbol{\theta}_0)$ and $l'_t(\boldsymbol{\theta}_0) = \partial l_t(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$.

Proof: This result uses the CLT for stationary and ergodic martingale difference sequences from Billingsley (1999).
We can start by arguing that $\{l'_t(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is a stationary and ergodic sequence by (Krengel, 1985, Proposition 4.3) since each of its elements is a continuous function of $\mathbf{g}_t(\boldsymbol{\theta})$ and $\partial \mathbf{g}_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ that are elements of stationary and ergodic sequences.
We can also argue that $\{l'_t(\boldsymbol{\theta}_0)\}_{t\in\mathbb{Z}}$ is a martingale difference sequence as $l_t(\boldsymbol{\theta}_0)$ is the conditional score of a correctly specified model. Our density meets the weak regularity conditions as it is a continuously differentiable function and its derivative with respect to $\boldsymbol{\theta}$ can be uniformly bounded by some constant in all its arguments. We need to show that the second moment of $l'_t(\boldsymbol{\theta})$ is bounded. Equivalently we can show that $||l_t(\boldsymbol{\theta}_0)||_2$ with $||\cdot||_n = (\mathbb{E}||\cdot||^n)^{1/n}$. Using the subadditivity of the norm for $n \geq 1$ we have,

$$\begin{aligned}
\left|\left|l'_t(\boldsymbol{\theta}_0)\right|\right|_2 &\leq \left|\left|\frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma})}{\partial \boldsymbol{\theta}}\right|\right|_2 + \left|\left|\frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma})}{\partial \varepsilon} \frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right|\right|_2 \\
&\leq \left|\left|\frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma})}{\partial \boldsymbol{\theta}}\right|\right|_2 + \left|\left|\frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma})}{\partial \varepsilon}\right|\right|_4 \cdot \left|\left|\frac{\partial \varepsilon_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right|\right|_4
\end{aligned}$$

We can see that the first term is bounded in (22). The second inequality comes instead from the Generalized Holder inequality and $\partial l_t(\varepsilon_t(\boldsymbol{\theta}_0), \boldsymbol{\gamma})/\partial \varepsilon$ is the expression in (**??**) and it is bounded. Then by the assumptions of *Theorem 2* we have that,

$$\left|\left|l'_t(\boldsymbol{\theta}_0)\right|\right|_2 < \infty$$

this finishes the proof of this part.

Lemma 4 *Under the assumptions 1- 4- 3 and if assumption 2 holds with $n \geq 4$ we have that,*

$$\sqrt{T} \sup_{\boldsymbol{\theta}\in\Theta} \left|\left|\hat{L}'_T(\boldsymbol{\theta}) - L'_T(\boldsymbol{\theta})\right|\right| \xrightarrow{a.s.} 0 \quad \textit{as } T \to \infty$$

Proof: We will show that,

$$\left|\left|\hat{l}'_t(\boldsymbol{\theta}) - l'_t(\boldsymbol{\theta})\right|\right|^{\boldsymbol{\theta}} \xrightarrow{e.a.s.} 0 \quad \text{as } t \to \infty.$$

As this implies that

$$\lim_{T\to\infty} \left|\left|\hat{L}'_T(\boldsymbol{\theta}) - L'_T(\boldsymbol{\theta})\right|\right|^{\boldsymbol{\theta}} \leq \sum_{t=1}^{T} \left|\left|\hat{l}'_t(\boldsymbol{\theta}) - l'_t(\boldsymbol{\theta})\right|\right|^{\boldsymbol{\theta}} < \infty \quad \text{a.s.}$$

And this implies our result. First using the subadditivity of the norm and relying on (17) we have that,

$$\left|\left|\hat{l}'_t(\boldsymbol{\theta}) - l'_t(\boldsymbol{\theta})\right|\right|^{\boldsymbol{\theta}} \leq \left|\left|\frac{\partial l_t(\hat{\varepsilon}_t(\boldsymbol{\theta}), \boldsymbol{\gamma})}{\partial \boldsymbol{\theta}} - \frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\gamma})}{\partial \boldsymbol{\theta}}\right|\right|^{\boldsymbol{\theta}} + \left|\left|\frac{\partial l_t(\hat{\varepsilon}_t(\boldsymbol{\theta}), \boldsymbol{\gamma})}{\partial \varepsilon} \frac{\partial \hat{\varepsilon}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\gamma})}{\partial \varepsilon} \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|\right|^{\boldsymbol{\theta}}.$$

This means that it is enough to show that the two terms on the right converge to zero. For the second term we rely on Corollary TA.16 of Blasques et al. (2021) according to which it is enough to show that,

$$\left|\left|\frac{\partial l_t(\hat{\varepsilon}_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial \varepsilon} - \frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial \varepsilon}\right|\right|^{\boldsymbol{\theta}} \xrightarrow{e.a.s.} 0 \quad \text{and} \quad \left|\left|\frac{\partial \hat{\varepsilon}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|\right|^{\boldsymbol{\theta}} \xrightarrow{e.a.s.} 0$$

if both $\partial l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})/\partial\varepsilon$ and $\partial\varepsilon_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ are stationary and ergodic and with a finite $\log^+||\cdot||$ moment. In our case we have that $\partial\varepsilon_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ from (20) is a continuous function of a stationary and ergodic sequence so it is stationary and ergodic by (Krengel, 1985, Proposition 4.3). It has also a bounded $||\cdot||$ as we can see from (20) and Proposition 2. The same reasoning holds for $\partial l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})/\partial\varepsilon$ (see (22) in the Appendix).
The convergence of $\partial\hat{\varepsilon}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ to $\partial\varepsilon_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ follows by the continuous mapping theorem and by Proposition 2.
The convergence of $\partial l_t(\hat{\varepsilon}_t(\boldsymbol{\theta}),\boldsymbol{\gamma})/\partial\varepsilon$ relies on mean value theorem,

$$\left|\left|\frac{\partial l_t(\hat{\varepsilon}_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial \varepsilon} - \frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial \varepsilon}\right|\right|^{\boldsymbol{\theta}} \leq \sup_{\boldsymbol{\theta}\in\Theta}\sup_{\varepsilon\in\mathbb{R}}\left|\frac{\partial^2 l(\varepsilon,\boldsymbol{\gamma})}{\partial\varepsilon^2}\right| \cdot |\hat{\varepsilon}_t(\boldsymbol{\theta}) - \varepsilon_t(\boldsymbol{\theta})|^{\boldsymbol{\theta}} \xrightarrow{e.a.s.} 0$$

which results from the fact that $\partial^2 l(\varepsilon,\boldsymbol{\gamma})\partial\varepsilon^2$ is bounded (see (22) in the Appendix) and by the convergence of $\hat{\mathbf{g}}_t(\boldsymbol{\theta})$ to $\mathbf{g}_t(\boldsymbol{\theta})$ e.a.s and the continuous mapping theorem.
The first term also follows from an application of the mean value theorem,

$$\left|\left|\frac{\partial l_t(\hat{\varepsilon}_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial \boldsymbol{\theta}} - \frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}),\boldsymbol{\gamma})}{\partial \boldsymbol{\theta}}\right|\right|^{\boldsymbol{\theta}} \leq \sup_{\boldsymbol{\theta}\in\Theta}\sup_{\varepsilon\in\mathbb{R}}\left|\frac{\partial^2 l(\varepsilon,\boldsymbol{\gamma})}{\partial\boldsymbol{\theta}\partial\varepsilon}\right| \cdot |\hat{\varepsilon}_t(\boldsymbol{\theta}) - \varepsilon_t(\boldsymbol{\theta})|^{\boldsymbol{\theta}} \xrightarrow{e.a.s.} 0$$

which results from (22) in the Appendix and by Proposition 1.

Lemma 5 *For any $n \geq 1$ such that $\mathbb{E}|\varepsilon_t|^n < \infty$ we have:*

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|^n < \infty \qquad \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\frac{\partial^2\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|\right|^n < \infty$$

Proof: We have,

$$\begin{aligned}\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\frac{\partial\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|^n &\leq \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\frac{\partial\Phi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta})\right|\right|^n + \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\Phi(\boldsymbol{\theta})\frac{\partial\mathbf{g}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|^n \\ &\leq K_1\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\mathbf{g}_t(\boldsymbol{\theta})\right|\right|^n + K_2\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\frac{\partial\mathbf{g}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|^n < \infty\end{aligned} \tag{14}$$

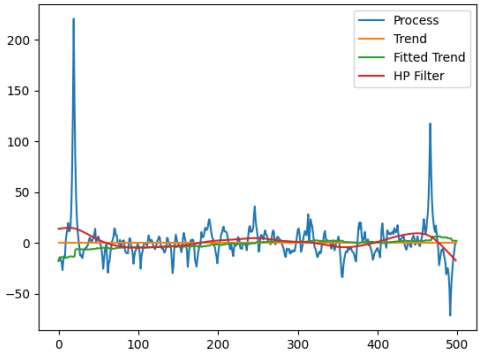
where $K_1 = \sup_{\boldsymbol{\theta}\in\Theta}||\partial\Phi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}||^n$ and $K_2 = \sup_{\boldsymbol{\theta}\in\Theta}||\Phi(\boldsymbol{\theta})||^n$ are bounded by $\boldsymbol{\theta}$ being compact set such that $\sigma > 0$ and the two other terms are less than infinity by Proposition 2. Then in a similar way we have,

$$\begin{aligned}\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\frac{\partial^2\varepsilon_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|\right|^n &\leq \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\frac{\partial^2\Phi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathbf{g}_t(\boldsymbol{\theta})\right|\right|^n + \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\frac{\partial\Phi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\mathbf{g}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|^n + \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\Phi(\boldsymbol{\theta})\frac{\partial^2\mathbf{g}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|\right|^n \\ &\leq K_1\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\mathbf{g}_t(\boldsymbol{\theta})\right|\right|^n + K_2\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\frac{\partial\mathbf{g}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\right|^n + K_3\left|\left|\frac{\partial^2\mathbf{g}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|\right|^n < \infty\end{aligned} \tag{15}$$

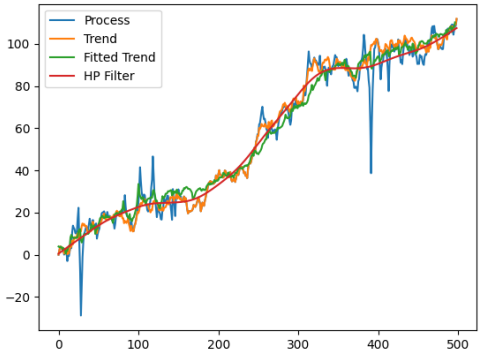
with the three terms being bounded by Proposition 2.

C Monte Carlo Simulation

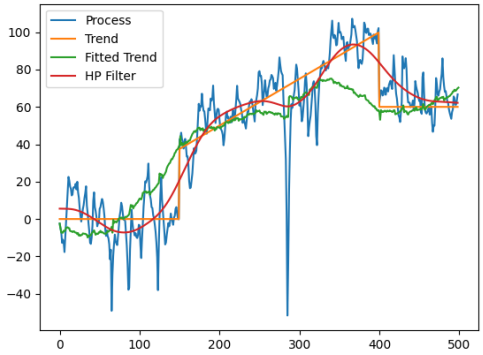
Figure 8: Simulated paths from the three data generating processes.



(a) MAR process without trend.



(b) MAR process plus random walk with drift.



(c) MAR process plus trend with breaks.

D Testing Procedure for Multimodal Predictive Densities

We use the Brier score, from Brier (1950), is computed as,

$$BS = \sum_{t=1}^{T} (p_t - o_t)^2$$

where p_t is the probability of our event and o_t is the realization of that event (1 if it happens, 0 otherwise). The range of this score is between 0 and 1. We now consider the multicategory Brier

score defined as,

$$BS = \frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{R}(p_{t,r} - o_{t,r})^2$$

where the r represents the different events and they must be such that $\sum_{r=1}^{R} p_{t,r} = 1$ for all t and $o_{t,r} = 1$ only for one r and it is 0 for the others. The range of this score is between 0 and 2. This multicategory score allows us to compare interval forecast. Since we are interested in prediction during a bubble (so we want to correctly address sharp increases and crashes) we consider as category movements that are within or outside the range of one standard deviation of a baseline Gaussian random walk. Our categories will then be,

$$p_{t,r} = \begin{cases} \mathbb{1}_{\Delta y_t < -\sigma_{rw}} & \text{if } r = 1 \\ \mathbb{1}_{|\Delta y_t| < \sigma_{rw}} & \text{if } r = 2 \\ \mathbb{1}_{\Delta y_t > \sigma_{rw}} & \text{if } r = 3 \end{cases} \tag{16}$$

With these scores we can create a Diebold Mariano test statistic. The test statistic for the multicategory Brier score will be,

$$d_t = \sum_{r=1}^{R}(p_{m,rt} - o_{m,rt})^2 - \sum_{r=1}^{R}(p_{i,rt} - o_{i,rt})^2$$

$$DM = \sqrt{T}\frac{\overline{d}}{\sigma_d}$$

where $\sigma_d = \sqrt{\hat{\gamma}(0) + 2\sum_{i=1}^{k} w_i\hat{\gamma}(i)}$, with k is of the same order as the square root of the test sample size and $w_i = 1 - i/k$.

E Additional derivations

Remember that $\boldsymbol{\theta} = (\omega, \alpha, \boldsymbol{\gamma})$ with $\boldsymbol{\gamma} = (\phi_1, ..., \phi_r, \psi_1, ..., \psi_s, \sigma, \nu)$ with $r+s = k$. The full derivatives of $L_T(\boldsymbol{\theta})$ are:

$$\frac{\partial l_t(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\gamma})}{\partial \boldsymbol{\theta}} = \frac{\partial l_t(\varepsilon, \boldsymbol{\gamma})}{\partial \boldsymbol{\theta}}\bigg|_{\varepsilon=\varepsilon_t(\boldsymbol{\theta})} + \frac{\partial l_t(\varepsilon, \boldsymbol{\gamma})}{\partial x}\bigg|_{\varepsilon=\varepsilon_t(\boldsymbol{\theta})} \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{17}$$

We now can consider that,

$$\frac{\partial l_t(\varepsilon, \boldsymbol{\gamma})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \frac{\partial l_t(\varepsilon, \boldsymbol{\gamma})}{\partial \nu} \end{pmatrix}$$

As the only parameter that actually appears in the likelihood is ν all the others are inside $\varepsilon_t(\boldsymbol{\theta})$. We also have,

$$\begin{aligned} \frac{\partial^2 l_t(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\gamma})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} =& \frac{\partial^2 l_t(\varepsilon, \boldsymbol{\gamma})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\bigg|_{\varepsilon=\varepsilon_t(\boldsymbol{\theta})} + \frac{\partial^2 l_t(x, \boldsymbol{\gamma})}{\partial \boldsymbol{\theta} \partial \varepsilon}\bigg|_{\varepsilon=\varepsilon_t(\boldsymbol{\theta})} \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{\partial^2 l_t(\varepsilon, \boldsymbol{\gamma})}{\partial \varepsilon^2}\bigg|_{\varepsilon=\varepsilon_t(\boldsymbol{\theta})} \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \\ &+ \frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial^2 l_t(\varepsilon, \boldsymbol{\gamma})}{\partial \varepsilon \partial \boldsymbol{\theta}}\bigg|_{\varepsilon=\varepsilon_t(\boldsymbol{\theta})} + \frac{\partial l_t(\varepsilon, \boldsymbol{\gamma})}{\partial \varepsilon}\bigg|_{\varepsilon=\varepsilon_t(\boldsymbol{\theta})} \frac{\partial^2 \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \end{aligned} \tag{18}$$

We now can consider that,

$$\frac{\partial^2 l_t(\varepsilon, \boldsymbol{\gamma})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \frac{\partial^2 l_t(\varepsilon, \boldsymbol{\gamma})}{\partial \nu \partial \nu'} & \end{pmatrix} \quad \text{and} \quad \frac{\partial^2 l_t(\varepsilon, \boldsymbol{\gamma})}{\partial \varepsilon \partial \boldsymbol{\theta}'} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{\partial^2 l(\varepsilon, \boldsymbol{\gamma})}{\partial \varepsilon \partial \nu} \end{pmatrix} \tag{19}$$

for the same reason as before.

We can now provide expressions for these derivatives dividing them in two groups:

A) The derivatives of the model for $l(\varepsilon, \boldsymbol{\gamma})$ and $s(\varepsilon, \boldsymbol{\gamma}) = \partial l(\varepsilon, \boldsymbol{\gamma})/\partial \varepsilon$. We know that:

$$l(\varepsilon, \boldsymbol{\gamma}) = -\frac{\log(\pi\nu)}{2} + \log\left(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\right) - \frac{\nu+1}{2}\log\left(1 + \frac{\varepsilon^2}{\nu}\right)$$

$$\frac{\partial l(\varepsilon, \boldsymbol{\gamma})}{\partial \varepsilon} = -(\nu+1)\frac{\varepsilon}{\nu+\varepsilon^2} \qquad \frac{\partial^2 l(\varepsilon, \boldsymbol{\gamma})}{\partial \varepsilon^2} = -(\nu+1)\frac{\nu-\varepsilon^2}{\left(\nu+\varepsilon^2\right)^2} \qquad \frac{\partial^2 l(\varepsilon, \boldsymbol{\gamma})}{\partial \nu \partial \varepsilon} = -\varepsilon\frac{1+\varepsilon^2}{\left(\nu+\varepsilon^2\right)^2}$$

$$\frac{\partial l(\varepsilon, \boldsymbol{\gamma})}{\partial \nu} = -\frac{1}{2\nu} + h'(\nu) + \frac{\nu+1}{2\nu}\frac{\varepsilon^2}{\nu+\varepsilon^2} - \frac{1}{2}\log\left(\frac{\nu+\varepsilon^2}{\nu}\right)$$

$$\frac{\partial^2 l(\varepsilon, \boldsymbol{\gamma})}{\partial \nu^2} = \frac{1}{2\nu^2} + h''(\nu) + \frac{\varepsilon^2 + \nu(\nu+2)}{2\nu^2}\frac{\varepsilon^2}{\left(\nu+\varepsilon^2\right)^2} + \frac{\varepsilon^2}{\varepsilon^2\nu + \nu^2}$$

B) The derivatives of the prediction error:

$$\varepsilon_t(\boldsymbol{\theta}) = \frac{\phi(L)\psi(L^{-1})}{\sigma} g_{t-s}(\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta})\mathbf{g}_t(\boldsymbol{\theta})$$

with $\mathbf{g}_t = \big[g_t, g_{t-1}, ..., g_{t-k+1}\big]'$ and $\Phi(\boldsymbol{\theta})$ the vector of coefficients from the MAR polynomial. Then:

$$\begin{aligned}
\frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \Phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{g}_t(\boldsymbol{\theta}) + \Phi(\boldsymbol{\theta}) \frac{\partial \mathbf{g}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
\frac{\partial^2 \varepsilon_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} &= \frac{\partial^2 \Phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathbf{g}_t(\boldsymbol{\theta}) + 2 \frac{\partial \Phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \mathbf{g}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \Phi(\boldsymbol{\theta}) \frac{\partial^2 \mathbf{g}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}
\end{aligned} \tag{20}$$

In the specific we have:

$$\begin{aligned}
\frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \omega} &= \Phi(\boldsymbol{\theta}) \frac{\partial \mathbf{g}_t(\boldsymbol{\theta})}{\partial \omega} \\
&= \Phi(\boldsymbol{\theta}) \big[\frac{\partial C(\boldsymbol{\theta})}{\partial \omega} + A(\boldsymbol{\theta}) \frac{\partial \mathbf{g}_{t-1}(\boldsymbol{\theta})}{\partial \omega} \big] \\
\frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \alpha} &= \Phi(\boldsymbol{\theta}) \frac{\partial \mathbf{g}_t(\boldsymbol{\theta})}{\partial \alpha} \\
&= \Phi(\boldsymbol{\theta}) \big[\frac{\partial A(\boldsymbol{\theta})}{\partial \alpha} \mathbf{g}_{t-1}(\boldsymbol{\theta}) + A(\boldsymbol{\theta}) \frac{\partial \mathbf{g}_{t-1}(\boldsymbol{\theta})}{\partial \alpha} \big] \\
\frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \gamma_i} &= \frac{\partial \Phi(\boldsymbol{\theta})}{\partial \gamma_i} \mathbf{g}_t(\boldsymbol{\theta}) + \Phi(\boldsymbol{\theta}) \frac{\partial \mathbf{g}_t(\boldsymbol{\theta})}{\partial \gamma_i} \quad \text{for } i = 1, ..., k+1 \\
&= \frac{\partial \Phi(\boldsymbol{\theta})}{\partial \gamma_i} \mathbf{g}_t(\boldsymbol{\theta}) + \Phi(\boldsymbol{\theta}) \big[\frac{\partial A(\boldsymbol{\theta})}{\partial \gamma_i} \mathbf{g}_{t-1}(\boldsymbol{\theta}) + A(\boldsymbol{\theta}) \frac{\partial \mathbf{g}_{t-1}(\boldsymbol{\theta})}{\partial \gamma_i} \big] \\
\frac{\partial \varepsilon_t(\boldsymbol{\theta})}{\partial \gamma_i} &= 0 \quad \text{for } i = k+2
\end{aligned} \tag{21}$$

We can write $\Phi(\boldsymbol{\theta}) = (\varphi_0, ..., \varphi_{k-1})$ where φ_i is the coefficient of z^{i-s} in $\frac{1}{\sigma}(1 - \phi_1 z - \phi_2 z^2 - ... - \phi_r z^r)(1 - \psi_1 z^{-1} - \psi_2 z^{-2} - ... - \psi_s z^{-s})$. Then:

$$\begin{aligned}
\varphi_i &= \sum_{j=1}^{i} \psi_{s-j+1} \phi_{i-j} \quad \text{for } i = 1, ..., s \\
\varphi_i &= \sum_{j=1}^{k-i+1} \psi_{j-1} \phi_{j+i-s-1} \quad \text{for } i = s+1, ..., k
\end{aligned}$$

Then:

$$\begin{aligned}
\frac{\partial \Phi(\boldsymbol{\theta})}{\partial \phi_1} &= \frac{1}{\sigma} \big[0, \psi_s ..., \psi_1, 1, 0, ..., 0\big] \\
\frac{\partial \Phi(\boldsymbol{\theta})}{\partial \phi_r} &= \frac{1}{\sigma} \big[0, 0, ..., 0, \psi_s ..., \psi_1, 1\big] \\
\frac{\partial \Phi(\boldsymbol{\theta})}{\partial \psi_s} &= \frac{1}{\sigma} \big[1, \phi_1, ..., \phi_r, 0, 0, ..., 0\big] \\
\frac{\partial \Phi(\boldsymbol{\theta})}{\partial \psi_1} &= \frac{1}{\sigma} \big[0, .., 0, 1, \phi_1, ..., \phi_r, , 0\big]
\end{aligned}$$

The derivative for $A(\boldsymbol{\theta})$ is very similar but it is in a matrix form and it has the α parameter as a scale in front of the first row.

Now let's show all these expressions are bounded, we have that:

A)

$$
\begin{aligned}
&\sup_{\boldsymbol{\gamma}} \left| \frac{\partial l(\varepsilon, \boldsymbol{\gamma})}{\partial \varepsilon} \right| \leq \sup_{\boldsymbol{\gamma}} \frac{\nu+1}{\nu+\varepsilon^2} \cdot |\varepsilon| \leq k|\varepsilon| \\
&\sup_{\boldsymbol{\gamma},\varepsilon} \left| \frac{\partial^2 l(\varepsilon, \boldsymbol{\gamma})}{\partial \varepsilon^2} \right| \leq \sup_{\boldsymbol{\gamma},\varepsilon} \frac{\nu+1}{\nu+\varepsilon^2} \cdot \left| \frac{\nu-\varepsilon^2}{\nu+\varepsilon^2} \right| \leq 2 \\
&\sup_{\boldsymbol{\gamma}} \left| \frac{\partial l(\varepsilon, \boldsymbol{\gamma})}{\partial \nu} \right| \leq \sup_{\gamma} \left| -\frac{1}{2\nu} + h'(\nu) - \frac{\nu+1}{2\nu} \frac{\varepsilon^2}{\nu+\varepsilon^2} - \frac{1}{2} \log\left(\frac{\nu+\varepsilon^2}{\nu} \right) \right| \\
&\qquad\qquad \leq \frac{3}{2} + |h'(\nu)| + k_2 |\varepsilon|^{\delta} \qquad (22) \\
&\sup_{\boldsymbol{\gamma}} \left| \frac{\partial^2 l(\varepsilon, \boldsymbol{\gamma})}{\partial \nu^2} \right| \leq \sup_{\boldsymbol{\gamma}} \left| \frac{1}{2\nu^2} + h''(\nu) + \frac{\varepsilon^2 + \nu(\nu+2)}{2\nu^2} \frac{\varepsilon^2}{\left(\nu+\varepsilon^2\right)^2} + \frac{\varepsilon^2}{\varepsilon^2\nu + \nu^2} \right| \\
&\qquad\qquad \leq 1 + |h''(\nu)| + \left| \frac{\varepsilon^4}{\left(\nu+\varepsilon^2\right)^2} \right| + \left| \frac{\varepsilon^4}{2\nu^2\left(\nu+\varepsilon^2\right)^2} \right| \\
&\sup_{\boldsymbol{\gamma}} \left| \frac{\partial^2 l(\varepsilon, \boldsymbol{\gamma})}{\partial \nu \partial \varepsilon} \right| \leq \sup_{\boldsymbol{\gamma}} \frac{|\varepsilon|}{\nu+\varepsilon^2} \frac{1+\varepsilon^2}{\nu+\varepsilon^2} \leq 1
\end{aligned}
$$