(When) Do Long Autoregressions Account for Changes in Parameters?*

Matei Demetrescu[a†] and Uwe Hassler[b]

[a]University of Bonn [b]Goethe University Frankfurt

December 20, 2013

Abstract

In order to construct forecasts for time series exhibiting structural changes, the paper examines long autoregressions, where the number of lagged endogenous regressors is growing with the sample size. First, we rigorously show that the OLS estimators are elementwise consistent for the true autoregressive coefficients even when a break in the mean is ignored, but that the sum of the estimators converges to unity under such misspecification. Thanks to this unit-root like behavior of the fitted model, the resulting conditional forecasts are consistent for the true values which take the break into account. As long as the dynamic structure is invariant over time, the robustness property of the forecasts holds true more generally, a) for a piecewise smoothly varying mean function, and b) under general autoregressive dynamics of possibly infinite order including stationary long memory. Second, under breaks in the dynamic structure, parameter estimators are asymptotically biased and, correspondingly, the forecasts from long autoregressions are biased themselves for the conditional mean. Simulations confirm the relevance of our asymptotic findings for finite samples.

Keywords Forecasting; breaks in parameters; smoothly varying means; long memory

JEL C22, C53

*__Acknowledgements:__ The authors would like to thank Rob Taylor for helpful comments.

†__Corresponding author:__ Hausdorff Center for Mathematics and Department of Economics, Bonn University, Adenauerallee 24-42, D-53113 Bonn, Germany. E-mail address: matei.demetrescu@uni-bonn.de.

1 Introduction

Dynamic modelling and forecasting of economic and financial time series under breaks in parameters is a topic of long history and with recent interest in econometrics; see e.g. the editorial by Timmermann and van Dijk (2013) to a special issue in the Journal of Econometrics or the recent review articles by Clements and Hendry (2011) and Rossi (2013). The present paper investigates the behaviour of long autoregressions estimated by ordinary least squares (OLS), where the number of lagged endogenous regressors is growing with the sample size, in the presence of ignored instability. The use of autoregressive (AR) models for forecasting purposes can be traced back at least to Akaike (1969); long autoregressions (LAR), or AR approximations, have been studied by without breaks by Berk (1974), Bhansali (1978), and Gonçalves and Kilian (2007) under more classical assumptions, while Poskitt (2007, 2008) extends the analysis to long memory and noninvertible processes. Wang, Bauwens, and Hsiao (2013) (WBH) open the floor for a discussion under breaks. They investigate experimentally several forecast strategies when the underlying process is fractionally integrated with breaks in the order of integration d and/or subject to means shifts. The clearly dominating strategy in samples of length $T = 200$ ignores eventual breaks and simply produces forecasts relying on a LAR.[1] Wang et al. (2013, Theorem 1) offer as theoretical explanation for their experimental evidence that a fractionally integrated process with break in the order of integration and/or in the mean has a stationary AR representation of infinite order characterized by fractional integration of a certain order d^* given as a convex combination of the orders before and after the break. This claim, however, is not correct, as we will see in Section 2; see (10) below.

Our paper sets the LAR on firm theoretical grounds with two contributions. First, we address the situation of an ignored mean shift under constant dynamics. If the process is autoregressive of finite order, the estimated coefficients from a LAR converge to the true parameters elementwise (Proposition 1 and Remark 1), while the sum of the LAR coefficients (the number thereof diverges with the sample size) converges to one (Remark 2). The latter is the reason why the LAR residuals behave as if the series had been differenced such that the shift in mean is effectively removed from the data. Consequently, the LAR forecast converges to the conditional mean as true forecast function (Corollary 1). In fact, this result holds under more general conditions. We allow for a piecewise (Hölder) continuous mean function with several breaks under AR dynamics of infinite order that may even display long memory. As long as the AR structure is invariant over time, the LAR forecast is unbiased for the conditional mean asymptotically (Proposition 2). To sum up our first contribution: under constant dynamics the proposal by WBH can

[1] AR approximations as alternatives to fractional modelling for forecasting purposes have been suggested as early as by Ray (1993) in the context of no breaks.

indeed be theoretically justified, and is very valuable since LAR results in robust forecasts irrespective of eventual mean shifts at unknown time and irrespective of a continuously varying mean function that does not have to be specified. Second, we turn to the case of breaks in the autoregressive parameters. Again, we obtain the limits of OLS-LAR (Proposition 3) that differ from the true parameters of the post-break period, such the the LAR forecasts are conditionally biased in the limit and miss the true forecast function (Remark 3). This shows that the favourable evidence presented in WBH under changes in persistence is a finite sample effect that does not carry over to larger samples. We present experimental evidence with growing sample sizes confirming our theoretical results.

The next section becomes precise on the model under breaks in parameters and clarifies where and why the theoretical underpinning of LAR by WBH is flawed. Section 3 deals, first, with the case of a mean shift under constant finite order dynamics, and, second, with a smoothly varying mean function subject to eventual breaks under dynamics of infinite order and long memory. Section 4 turns to instability in the dynamic structure. In the fifth section, our asymptotic results are illustrated experimentally for a large variety of finite sample sizes. Concluding remarks are offered in the last section, and the mathematical proofs are collected in the Appendix.

Finally a word on notation: $\|\cdot\|$ is the Euclidean vector norm and the corresponding induced matrix norm, $\lfloor x \rfloor$ denotes the integer part of a positive number x, probabilistic Landau symbols $o_p(\cdot)$ and $O_p(\cdot)$ have their usual meanings, while $\xrightarrow{p}$ stands for convergence in probability as the sample size T goes off to infinity.

2 Model

We assume to observe T observations of a univariate process with changing parameters over time:

$$y_t = m_t + \begin{cases} x_t^{(1)}, & t = 1, 2, \ldots, T_1 = \lfloor \tau T \rfloor \\ x_t^{(2)}, & t = T_1 + 1, T_1 + 2, \ldots, T \end{cases} . \tag{1}$$

In the most general case, $\{m_t\}$ is only assumed to be piecewise Hölder continuous; see Assumption 4 below for details. The leading case, however, will be the one by WBH where $\{m_t\}$ simply captures a mean-shift,

$$y_t = \begin{cases} \mu_1 + x_t^{(1)}, & t = 1, 2, \ldots, T_1 = \lfloor \tau T \rfloor \\ \mu_2 + x_t^{(2)}, & t = T_1 + 1, T_1 + 2, \ldots, T \end{cases} . \tag{2}$$

At the same time the dynamic structure may be subject to a change,

$$x_t^{(r)} = A_r^{-1}(L)\varepsilon_t = \sum_{j=0}^{\infty} c_j^{(r)} \varepsilon_{t-j}\,, \quad \sum_{j=0}^{\infty} \left(c_j^{(r)}\right)^2 < \infty\,, \tag{3}$$

with break fraction $\tau \in [0,1]$. To distinguish the two regimes we use superscripts or subscripts $r \in \{1,2\}$. With the autoregressive polynomials A_r in the usual lag operator L, we may summarize (2) and (3) as

$$\begin{array}{l} A_1(L)y_t = A_1(1)\mu_1 + \varepsilon_t\,, \quad t \le T_1 \\ A_2(L)y_t = A_2(1)\mu_2 + \varepsilon_t\,, \quad t > T_1 \end{array} \quad , \quad A_r(L) = 1 - \sum_{j=1}^{\infty} a_j^{(r)} L^j\,,\ \sum_{j=1}^{\infty} \left(a_j^{(r)}\right)^2 < \infty\,. \tag{4}$$

In general, for the process to have a bounded mean, it must hold $A_r(1) = 1 - \sum_{j=1}^{\infty} a_j^{(r)} < \infty$ if $\mu_r \neq 0$. Further, we maintain the assumption that the innovations form a sequence of identically and independently distributed (*iid*) errors.

Assumption 1 *The sequence* $\{\varepsilon_t\}$ *is iid* $(0, \sigma^2)$ $\forall t \in \mathbb{Z}$ *with finite 4th order moments.*

It is noteworthy that $\{y_t\}$ has no stationary autoregressive representation neither in the case of a mean shift nor in the presence of a break in the dynamics. Filtering the process with any $A_*(L)$ results in

$$A_*(L)y_t = A_*(1)\mu_r + A_*(L)A_r^{-1}(L)\varepsilon_t\,, \quad r \in \{1,2\}\,. \tag{5}$$

Hence, $A_*(L)y_t = m + \varepsilon_t$ for all t *if and only if* $A_1 = A_2 = A_*$ and $\mu_1 = \mu_2$.

For forecasting purposes, let us consider a LAR of y_t (estimated by OLS) ignoring eventual breaks in parameters,

$$y_t = \hat{m} + \sum_{j=1}^{h_T} \hat{a}_{j,h_T} y_{t-j} + \hat{\varepsilon}_t = \hat{m} + \hat{\boldsymbol{a}}_{h_T}' \boldsymbol{y}_{t-h_T} + \hat{\varepsilon}_t, \tag{6}$$

where $\boldsymbol{y}_{t-h_T} = (y_{t-1}, \ldots, y_{t-h_T})'$ and $\hat{\boldsymbol{a}}_{h_T}$ is the vector of OLS estimators. Let $\boldsymbol{a}_{h_T} = (a_1, \ldots, a_{h_T})'$ denote the vector of the first h_T true parameters. In this setup, h_T is a function of T such that $h_T \to \infty$ and $h_T/T \to 0$ at suitable rates. In practice one would either use some deterministic function of T or determine h_T in a data-driven manner, say using information criteria. The implied one step ahead forecast function is then

$$\hat{y}_T(1) = \hat{m} + \hat{\boldsymbol{a}}_{h_T}' \boldsymbol{y}_{T+1-h_T} = \hat{m} + \sum_{j=1}^{h_T} \hat{a}_{j,h_T} y_{T+1-j},$$

while the true forecast function is given by

$$y_T(1) = \mathrm{E}\left(y_{T+1}|y_T, y_{T-1}, \ldots\right).$$

The use of long autoregressions like in (6) has been advocated by WBH assuming fractionally integrated noise in (4), i.e.

$$A_r(L) = (1-L)^{d_r} = 1 - \sum_{j=1}^{\infty} a_{j,d}^{(r)} L^j\,,\ |d_r| < 0.5\,, \tag{7}$$

with $a_{j,d}^{(r)} = -\pi_{j,d}^{(r)}$ and

$$\begin{aligned} \pi_{j,d}^{(r)} &= \frac{j-1-d_r}{j}\,\pi_{j-1,d}^{(r)}\,, \quad j \geq 1\,,\ \pi_{0,d}^{(r)} = 1 \\ &\sim \frac{j^{-d_r-1}}{\Gamma(-d_r)}\,, \quad j \to \infty\,. \end{aligned}$$

For negative d_r the binomial expansion of $(1-L)^{d_r}$ is not summable, so that we now assume $d_r > 0$ for $A_r(1)\mu_r$ in (4) to be defined. For $d_r > 0$, however, one has $1 - \sum_{j=1}^{\infty} a_{j,d}^{(r)} = 0$, so that we may drop the means altogether. Therefore, we assume for now that $\mu_1 = \mu_2 = 0$. Wang et al. (2013, Lemma 1) state that $\{y_t\}$ from (4) with (7) has a representation as a fractionally integrated process without break,

$$(1-L)^{d^*} y_t = \varepsilon_t\,, \quad d^* = \lambda d_1 + (1-\lambda) d_2\,,\ \lambda \in [0,1]\,, \tag{8}$$

where the apparent order of fractional integration d^* is a convex combination of d_1 and d_2, such that

$$(1-L)^{d^*} y_t = y_t - \sum_{j=1}^{\infty} a_j^{(*)} y_{t-j} = \varepsilon_t\,, \tag{9}$$

with the autoregressive coefficients $a_j^{(*)}$ taken from the expansion of $(1-L)^{d^*}$. This statement is not correct as can be been from (5). In fact, differencing y_t from (4) under (7) results under $\mu_1 = \mu_2 = 0$ in

$$(1-L)^{d^*} y_t = \begin{cases} (1-L)^{d^*-d_1}\varepsilon_t \sim I(d_1 - d^*) \\ (1-L)^{d^*-d_2}\varepsilon_t \sim I(d_2 - d^*) \end{cases}. \tag{10}$$

Hence, the process y_t is $I(d^*)$ only under $d_1 = d_2 = d^*$ (no break), which corrects the claim by Wang et al. (2013, Lemma 1). In all other cases there exists no white noise sequence that, upon filtering with $(1-L)^{-d^*}$, recovers the y_t series with breaks.

Let us briefly understand why WBH are mislead to claim (8). Their argument builds on establishing the following variance behaviour:

$$\text{Var}\left(\sum_{t=1}^{T} y_t\right) = O\left(T^{2d^*+1}\right) . \tag{11}$$

Such a variance behaviour is implied by fractional integration of order d^*; but there is no equivalence between (11) and fractional integration of order d^*. Indeed, many processes that satisfy (11) but are not $I(d^*)$ have been suggested under the label of "spurious long memory". See amongst others Engle and Smith (1999), Diebold and Inoue (2001), and Granger and Hyung (2004); the first paper we are aware of that addressed the potential confusion of long memory ("Hurst phenomenon") and mean shifts is by (Klemeš, 1974, p. 675): "It is shown that the Hurst phenomenon is not necessarily an indicator of infinite memory of a process. It can also be caused by nonstationarity in the mean [...]". Recently, the variance behaviour in (11) has been related to the concept of "summability of order d^*" introduced by Berenguer-Rico and Gonzalo (2014) of which fractional integration is a special case.

Assuming that Wang et al. (2013, Lemma 1) is correct, Wang et al. (2013) try to provide grounds for a forecasting strategy building on LAR ignoring eventual breaks in mean and/or memory. Given the autoregressive representation (9), Wang et al. (2013, Theorem 1) argue that (6) results in consistent estimators converging to $a_j^{(*)}$ at a rate depending on h_T. In the model with break, however, such a convergence cannot take place, simply because the assumed AR representation in (9) does not exist. Nevertheless, WBH provide very promising experimental evidence on LAR as a forecast device. Since their theoretical justification is flawed, one question comes in naturally: in what situation do LAR actually account for breaks in parameters?

3 Changes in the mean

In this section we focus on the case where $A_1(L) = A_2(L) = A(L)$. With respect to the mean function we begin with the special case of (2) and then move on to the more general model (1). Similarly, the first subsection is restricted to the situation of a finite order AR(p) process rendering itself to simpler interpretation, while the second subsection is reserved for AR(∞) and long memory.

3.1 AR(p) with break in mean

For polynomials $A_1(L) = A_2(L)$ constant over time, the model in (2) reduces to a stationary process except for the mean shift,

$$y_t = m_t + x_t\,, \tag{12}$$

where the deterministic mean function m_t exhibits a jump:

$$m_t = \begin{cases} \mu_1 = -\left(1-\tau\right)\left(m_2 - m_1\right)\,, & t \leq \tau T \\ \mu_2 = \tau\left(m_2 - m_1\right)\,, & t > \tau T \end{cases}. \tag{13}$$

To simplify matters, we assume a demeaned structural break m_t. Hence, we do not have to allow for an intercept in the long autoregression (6) without loss of generality.

In this subsection, the assumptions on the stochastic component are as follows.

Assumption 2 *The process $\{x_t\}$ is autoregressive of finite order p given by $A\left(L\right)x_t = x_t - \sum_{j=1}^{p} a_j x_{t-j} = \varepsilon_t\ \forall t \in \mathbb{Z}$ where $\{\varepsilon_t\}$ is from Assumption 1, and $A\left(z\right)$ has all roots outside the unit circle. Let $\Sigma_{h_T} = \mathrm{Cov}\left(\boldsymbol{x}_{t-h_T}\right)$ denote the h_Tth order covariance matrix of $\{x_t\}$, and $\Gamma_{h_T} = \mathrm{E}(\boldsymbol{x}_{t-h_T}x_t)$ where $\boldsymbol{x}_{t-h_T} = (x_{t-1},\ldots,x_{t-h_T})'$.*

Following Clements and Hendry (2006), the occurrence of a structural break is not only a matter of the data generating process but also of the model employed. If one manages to define a step dummy variable D_t indicating the break point correctly, and fits $y_t = \mu_1 + \mu_2 D_t + x_t$ to the data from (12), the extended model with parameters μ_1 and μ_2 does not suffer from a structural break. Omitting the dummy variable D_t, however, typically results in an omitted variable bias. We will now show why and how the long autoregression overcomes this omitted variable bias.

For the data generating process (DGP) in Assumption 2, it is known that the eigenvalues of Σ_{h_T} and $\Sigma_{h_T}^{-1}$ are bounded and bounded away from zero, such that $\|\Sigma_{h_T}\| = O\left(1\right)$ and $\left\|\Sigma_{h_T}^{-1}\right\| = O\left(1\right)$. See the Fundamental Theorem of Grenander and Szegö given for instance in Brockwell and Davis (1991, Prop. 4.5.3). This will be used to establish the following result.

Proposition 1 *Let $\boldsymbol{a}_{h_T} = (a_1,\ldots,a_p,0,\ldots,0)' \in \mathbb{R}^{h_T}$ denote the vector of true parameters, and define*

$$\tilde{\boldsymbol{a}}_{h_T} = \boldsymbol{a}_{h_T} + \frac{\bar{\mu}^2}{1 + \bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\Sigma_{h_T}^{-1}\boldsymbol{\iota}\left(1 - \boldsymbol{\iota}'\boldsymbol{a}_{h_T}\right)$$

where $\boldsymbol{\iota}$ is an h_T-vector of ones and

$$\bar{\mu}^2 = \tau(1-\tau)(\mu_2 - \mu_1)^2.$$

If $h_T^{-1} + h_T T^{-\kappa} \to 0$ for some $\kappa \in \left(0, \frac{1}{3}\right)$, it holds under (12) with (13) and Assumption 2 that

$$\|\hat{\boldsymbol{a}}_{h_T} - \tilde{\boldsymbol{a}}_{h_T}\| = o_p\left(h_T^{-0.5}\right)$$

as $T \to \infty$.

Proof: *See the Appendix.*

The sequence $\tilde{\boldsymbol{a}}_{h_T}$ forms a triangular array, and $\widetilde{a}_{j,h_T}$ changes for fixed j with the sample size. How close $\tilde{\boldsymbol{a}}_{h_T}$ and $\boldsymbol{a}_{h_T}$ are depends on the magnitude and the timing of the jump through $\bar{\mu}^2$, where the effect of the break point is symmetric about 0.5. E.g. for the special case where $\{x_t\}$ is white noise we have for large h_T

$$\tilde{\boldsymbol{a}}_{h_T} = \frac{\frac{\bar{\mu}^2}{\sigma^2}}{1 + \frac{\bar{\mu}^2}{\sigma^2} h_T} \boldsymbol{\iota} \approx \frac{1}{h_T} \boldsymbol{\iota}.$$

This nicely illustrates the first-order limiting properties of $\hat{\boldsymbol{a}}_{h_T}$ discussed the the following two remarks.

Remark 1 The proposition implies *elementwise* convergence of the LAR OLS estimators, $\hat{a}_{j,h_T} \xrightarrow{p} a_j$ for each $j \leq p$ and $\hat{a}_{j,h_T} \xrightarrow{p} 0$ for each $p < j \leq h_T$ even when ignoring breaks in the mean. This is because $\boldsymbol{\iota}' \Sigma_{h_T}^{-1} \boldsymbol{\iota} \to \infty$ as $h_T \to \infty$ and the row sums of $\Sigma_{h_T}^{-1}$ are bounded.

Hence the dynamics of the process are in a sense recovered in spite of not accounting for breaks in the mean. But this is only half the story. The remark does not explain why the neglected mean shift would not affect the forecasts, which are after all centered at the post-break mean. The following remark sheds light on this issue.

Remark 2 Because of $\boldsymbol{\iota}' \Sigma_{h_T}^{-1} \boldsymbol{\iota} \to \infty$ one obtains

$$\begin{aligned} \sum_{j=1}^{h_T} \tilde{a}_{j,h_T} &= \boldsymbol{\iota}' \boldsymbol{a}_{h_T} + \frac{\bar{\mu}^2 \boldsymbol{\iota}' \Sigma_{h_T}^{-1} \boldsymbol{\iota}}{1 + \bar{\mu}^2 \boldsymbol{\iota}' \Sigma_{h_T}^{-1} \boldsymbol{\iota}} \left(1 - \boldsymbol{\iota}' \boldsymbol{a}_{h_T}\right) = \boldsymbol{\iota}' \boldsymbol{a}_{h_T} + (1 + o(1))\left(1 - \boldsymbol{\iota}' \boldsymbol{a}_{h_T}\right) \\ &\to 1 \end{aligned}$$

since $1 - \boldsymbol{\iota}' \boldsymbol{a}_{h_T}$ is bounded thanks to the stability of $\{x_t\}$. In other words: the fitted LAR seemingly has a unit root in that the sum of its coefficients is unity in the limit, which washes out the change in mean when forecasting by effectively differencing it away.

We now take a more rigorous look at the long autoregressive forecast function and show it to be consistent for the true one, given by

$$y_T(1) = \mathrm{E}\left(y_{T+1}|y_T, y_{T-1}, \ldots\right) = \mu_2 + \boldsymbol{x}'_{T+1-h_T}\boldsymbol{a}_{h_T}\,.$$

We have the following result.

Corollary 1 *Under the assumptions of Proposition 1 it holds*

$$\widehat{y}_T(1) = y_T(1) + o_p(1)$$

as $T \to \infty$.

Proof: *See the Appendix.*

3.2 Extensions

We now extend the model (12) from Proposition 1 in two directions. First, we step beyond the AR process of finite order from Assumption 2 and allow for AR(∞) with or without long memory. Second, we replace (13) and consider a more general mean function as indicated in (1). We will find that results analogous to Proposition 1 with Remark 2 hold true under such much more general conditions, and the robustness property from Corollary 1 carries over.[2]

Assumption 3 *For* $0 \leq d < 1/2$ *the stationary process* $\{x_t\}$ *is given by* $(1-L)^d x_t = B(L)\varepsilon_t$ *where* $\{\varepsilon_t\}$ *obeys Assumption 1. The coefficients of* $B(L) = \sum_{j=0}^{\infty} b_j L^j$ *with* $b_0 = 1$ *satisfy* $\sum_{j=0}^{\infty} |b_j| < \infty$, $\sum_{j=0}^{\infty} b_j \neq 0$, *and* $j^{1-d}b_j \to 0$ *as* $j \to \infty$. *Further,* $\Sigma_{h_T}^{-1}\Gamma_{h_T}$ *denotes the coefficients of the best linear predictor of* x_t *given* $\boldsymbol{x}_{t-h_T} = (x_{t-1}, \ldots, x_{t-h_T})'$.

The stationary process $\{x_t\}$ has a Wold representation where the coefficients are given by convolution: $x_t = (1-L)^{-d} B(L)\varepsilon_t$. The usual expansion of $(1-L)^{-d}$ results in coefficients with the decay rate j^{d-1} that is characteristic for fractional integration. For the long memory case $d > 0$, we adopt from Hassler and Kokoszka (2010, Prop. 2.1) the assumption $j^{1-d}b_j \to 0$ on $B(L)$, which is necessary and sufficient for the hyperbolic rate j^{d-1} to carry over from the filter $(1-L)^{-d}$ to the Wold coefficients of $\{x_t\}$. For $d = 0$, $\{x_t\}$ is simply integrated of order 0.

[2]The robustness of methods when analyzing long memory under trends has been investigated previously e.g. by Bhattacharya et al. (1983) and Giraitis et al. (2001).

Now, we turn to the mean process. There is in fact no a priori reason to assume just one single break in (13); we may allow, more generally, for several such discontinuities. Moreover, $\{m_t\}$ does not have to be constant between two breaks; we only require continuity, more precisely only Hölder continuity of some order α. For a function $\nu(\cdot)$ on $[0,1]$ we hence assume

$$\sup_{0\leq s<t\leq 1} \frac{|\nu(t)-\nu(s)|}{|t-s|^{\alpha}} < \infty\,, \quad 0<\alpha\leq 1\,.$$

Assumption 4 *The mean process $\{m_t\}$ is given by $m_t = \nu(t/T)$, where $\nu(\cdot)$ is piecewise Hölder continuous such that the discontinuities are interior points of $[0,1]$. Further, we assume $\int_0^1 \nu(s)\,\mathrm{d}s = 0$, and denote $\bar{\mu}^2 = \int_0^1 \nu^2(s)\,\mathrm{d}s$.*

This assumption encompasses, in addition to sudden breaks, a slowly evolving trend or a random level model. For instance, a Wiener process possesses the pathwise property from Assumption 4 for any $0<\alpha<1/2$ so $\nu(s)\equiv W(s)$ is allowed for. The simplifying condition $\int \nu(s)\,\mathrm{d}s = 0$ assumes that the process is demeaned. Hence, we consider again a LAR without an intercept without loss of generality.

Proposition 2 *Consider $\{y_t\}$ from (12) with $\{x_t\}$ from Assumption 3, and $\{m_t\}$ satisfies Assumption 4 with $1/4<\alpha\leq 1$. Then, for h_T such that $h_T^{-1}+h_T T^{-\kappa}\to 0$ for some $0<\kappa<\min\left\{\frac{\alpha}{\alpha+1.5+2d};\frac{1-2d}{3+4d}\right\}$, it follows that*

$$\|\hat{\boldsymbol{a}}_{h_T} - \tilde{\boldsymbol{a}}_{h_T}\| = o_p\left(h_T^{-0.5}\right),$$

as $T\to\infty$, where

$$\tilde{\boldsymbol{a}}_{h_T} = \Sigma_{h_T}^{-1}\Gamma_{h_T} + \frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\Sigma_{h_T}^{-1}\boldsymbol{\iota}\left(1-\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\Gamma_{h_T}\right)$$

with $\Sigma_{h_T}^{-1}\Gamma_{h_T}$ and $\bar{\mu}^2$ from Assumption 3 and 4, respectively.

Moreover, Corollary 1 continues to hold.

Proof: *See the Appendix.*

The choice of κ is more limited than in Proposition 1. On the one hand, the presence of long memory imposes $\kappa<\frac{1-2d}{3+4d}$. This is even stricter than the rate derived by Poskitt (2007), and is due to the presence of a piecewise smoothly varying mean function not accounted for in the LAR. The intuition behind the rate reduction is that otherwise vanishing terms cumulate over $\hat{a}_{j,h_T}$ such that h_T must be reduced in order to maintain the first-order limiting behavior derived in Proposition 2. The additional restriction $\kappa<$

$\frac{\alpha}{\alpha+1.5+2d}$ is due to the smoothness (or rather roughness) condition on the mean function ν and has essentially the same interpretation. It is not binding, for instance, when ν satisfies a Lipschitz condition, i.e. when $\alpha = 1$. The additional restriction for κ depends on the local properties of ν which may not be easily estimated, but one can always pick it conservatively as $\frac{\alpha_{\min}}{2.5+2d}$ for some $\alpha_{\min} > 1/4$ that one is prepared to accept. The "worst-case" scenario would be $\kappa < 1/11$ for a lower bound of $1/4$ for α and a conservative $d = 1/2$. But when d is close to $1/2$, it is rather $\frac{1-2d}{3+4d}$ that is binding: for $d > 4/13$, $\frac{1-2d}{3+4d} < 1/11$. A logarithmic rate for h_T satisfies both.

4 Breaks in the autoregressive coefficients

As a special case of (2) we now consider the situation of breaks in the dynamic structure,

$$y_t = \begin{cases} x_t^{(1)} = A_1^{-1}(L)\,\varepsilon_t, & t \leq \tau T \\ x_t^{(2)} = A_2^{-1}(L)\,\varepsilon_t, & t > \tau T \end{cases}, \tag{14}$$

under the simplifying assumption of a constant mean equal to zero. Since it will turn out that in this simplest case the LAR does not yield a valid forecast function, this will be all the more true for more complicated structures. For the same reason we assume the AR polynomials to be of finite order and need not examine the AR(∞) case.

The true forecast function is based on A_2, i.e.

$$y_T(1) = \sum_{j=1}^{p} a_j^{(2)} y_{T+1-j}\,.$$

Again, the break is ignored and a long autoregression of order h_T is fitted, intending to use it for forecasting:

$$\hat{y}_T(1) = \sum_{j=1}^{h_T} \hat{a}_{j,h_T} y_{T+1-j}.$$

The process from (14) is nonstationary, and does not have a Wold representation. Still, we may examine the first-order asymptotics of the OLS estimators like before, in order to subsequently analyze the forecast function.

Proposition 3 *Let $\{y_t\}$ be from (14) and $x_t^{(r)}$, $r = 1, 2$, satisfy Assumption 2 each, with true parameter vectors $\boldsymbol{a}_{h_T}^{(r)} = \left(a_1^{(r)}, \ldots, a_p^{(r)}, 0, \ldots, 0\right)'$. Define*

$$\bar{\boldsymbol{a}}_{h_T} = \left(I_{h_T} + \frac{1-\tau}{\tau}\Sigma_{h_T,1}^{-1}\Sigma_{h_T,2}\right)^{-1}\left(\boldsymbol{a}_{h_T}^{(1)} + \frac{1-\tau}{\tau}\Sigma_{h_T,1}^{-1}\Sigma_{h_T,2}\boldsymbol{a}_{h_T}^{(2)}\right).$$

If $h_T^{-1} + h_T T^{-\kappa} \to 0$ *for some* $\kappa \in \left(0, \frac{1}{3}\right)$*, it holds that*

$$\|\hat{\boldsymbol{a}}_{h_T} - \bar{\boldsymbol{a}}_{h_T}\| = o_p\left(h_T^{-0.5}\right)$$

as $T \to \infty$*.*

Proof: *Analogous to the proof of Proposition 1 and omitted.*

Unlike the case of a break in the mean we no longer have elementwise convergence to $\boldsymbol{a}_{h_T}^{(2)}$. We stress this fact and the consequences for forecasting in the following remark.

Remark 3 Consider for simplicity the case $p = 1$ where in the first regime $a_1^{(1)} \neq 0$, while the postbreak regime is characterized by white noise, i.e. $a_1^{(2)} = 0$. Then

$$\bar{\boldsymbol{a}}_{h_T} = \left(I_{h_T} + \frac{1-\tau}{\tau}\Sigma_{h_T,1}^{-1}\right)^{-1} \boldsymbol{a}_{h_T}^{(1)}$$

where $\boldsymbol{a}_{h_T}^{(1)} = \left(a_1^{(1)}, 0, \ldots, 0\right)'$ and $\Sigma_{h_T,1}^{-1}$ is correspondingly a positive definite band matrix, so $I_{h_T} + \frac{1-\tau}{\tau}\Sigma_{1,h_T}^{-1}$ has eigenvalues bounded and bounded away from zero. Thus $\bar{\boldsymbol{a}}_{h_T}$ must be nonzero since it equals $a_1^{(1)}$ times the first row of $\left(I_{h_T} + \frac{1-\tau}{\tau}\Sigma_{h_T,1}^{-1}\right)^{-1}$; and since this inverse exists, its first row is nonzero. But the required limit for a correct forecast is, at the end of the sample, the true vector $\boldsymbol{a}_{h_T}^{(2)} = (0, 0, \ldots, 0)'$. This shows that $\hat{y}_T(1)$ is biased for the conditional forecast $y_T(1)$ even asymptotically.

It is interesting to add some intuition to Proposition 3. The first regime of the sample (which should be irrelevant for forecasting at the end of the second one) has an effect on the estimator, weighted by τ. This parallels the situation where some process of interest has no break but is superimposed by disturbances with own dynamics. To become precise, let

$$y_t = (1-\tau)\, A_2^{-1}\varepsilon_t^{(2)} + \tau A_1^{-1}\varepsilon_t^{(1)}\,,$$

where $\{\varepsilon_t^{(1)}\}$ is independent of $\{\varepsilon_t^{(2)}\}$. Then for any fixed autoregressive order $\bar{p}$ the limit of the OLS autoregressive estimators for y_t is given by

$$\left(\tau\Sigma_{\bar{p},1} + (1-\tau)\,\Sigma_{\bar{p},2}\right)^{-1}\left(\tau\Gamma_{\bar{p},1} + (1-\tau)\,\Gamma_{\bar{p},2}\right)\,.$$

Of course one encounters the typical errors in variables effect. Note that the limit under errors in variables is essentially the same expression as the one derived in Proposition 3. An analogous result can be shown to hold when the order is $\bar{p} = h_T \to \infty$. Hence, ignoring changes in the dynamics when running a LAR amounts to estimating under measurement errors, and forecasting the signal component using the estimated dynamics

of signal *with* noise. Therefore, inconsistency may not come as a surprise, and there is no hope to construct unbiased conditional forecasts without accounting for the breaks.

5 Simulation evidence

In order to assess the finite-sample relevance of our limiting results, we conduct a Monte Carlo analysis examining four particular situations. First, $\{y_t\}$ exhibits a break in the mean but has otherwise homogenous AR(1) dynamics. Second, $\{y_t\}$ is fractionally integrated noise of order d having a break in the mean. Third, $\{y_t\}$ is a zero-mean AR(1) process with a break in the autoregressive parameter, and fourth, $\{y_t\}$ has a constant mean zero with fractionally integrated noise subject to a break in the integration order d.

For all four scenarios we examine series of length $T \in \{50, 100, 200, 500, 1000, 2000\}$ with a burn-in period of 100 observations that are discarded. The shocks ε_t are standard normal independent white noise, and the results rely on 25000 replications for each parameter constellation. The lag length h_T of the LAR is chosen by Akaike's information criterion, AIC, with a maximum order given by $\lfloor 12(T/100)^{0.25} \rfloor$. We report a) the in-sample residual variance averaged over the 25000 replications, and b) the variance over 25000 replications of the difference between the fitted forecast function $\hat{y}_T(1)$ and the true forecast function $y_T(1)$. We report both, since the residual variance averages over the entire series, whereas the difference between the forecast functions, although only relevant at the end of the sample, quantifies the optimality loss of the forecast, and this is the relevant figure for practitioners.

The simulated data generating processes are for the four scenarios as follows.

1. For the AR(1) process with a break in the mean we simulate with an autoregressive parameter $a_1 \in \{0, 0.1, 0.3, 0.5, 0.7\}$. The break fraction is taken to be $\tau = 0.5$, and the magnitude of the break is either $\mu_2 - \mu_1 = 0.2$ or $\mu_2 - \mu_1 = 1$.

2. For Scenario 2, we use the same setup as in Scenario 1 for the discontinuity in the mean function, but $\{y_t\}$ has fractionally integrated noise with $d \in \{0, 0.1, 0.2, 0.3, 0.4\}$ for the purely stochastic component.

3. Third, for the AR(1) process with a break in the autoregressive parameter we have $\tau = 0.3$ or $\tau = 0.7$; the autoregressive parameter breaks from $a_1^{(1)} \in \{0, 0.1, 0.3, 0.5, 0.7\}$ for the pre-break sample to independent white noise ($a_1^{(2)} = 0$) for the post-break period.

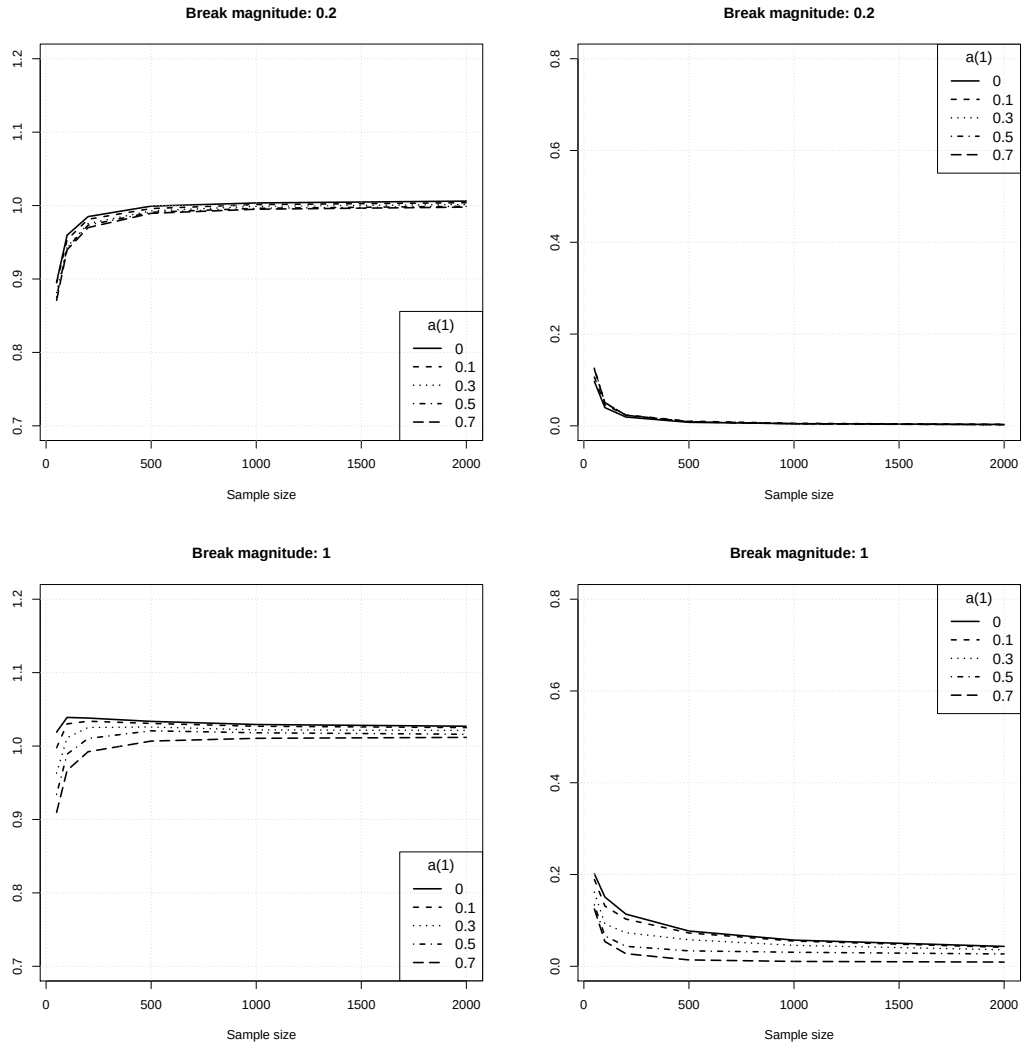

Figure 1: Average of residual variance (left), and variance of difference between true and LAR forecast (right) for AR(1) processes with break in mean

4. Finally, for the fractionally integrated process with break in d but not in the mean we have the setup analogous to that of Scenario 3, with $d_1 \in \{0, 0.1, 0.2, 0.3, 0.4\}$ before the break and independent white noise ($d_2 = 0$) thereafter.

The results for the four scenarios are as follows.

1. Scenario 1; see Figure 1: For a small break in mean ($\mu_2 - \mu_1 = 0.2$), the in-sample residual variance is close to the theoretical one ($\sigma^2 = 1$), at least for larger sample sizes, while at the same time the Monte Carlo variance of $\hat{y}_T(1) - y_T(1)$ is close to zero, which illustrates Proposition 1 and Corollary 1, respectively. For a larger break in mean ($\mu_2 - \mu_1 = 1$) the correspondence between the experimental and the asymptotic values is not quite so close, and it takes some larger sample to kick in, especially for the variance of the differences between the forecasts. Interestingly, the

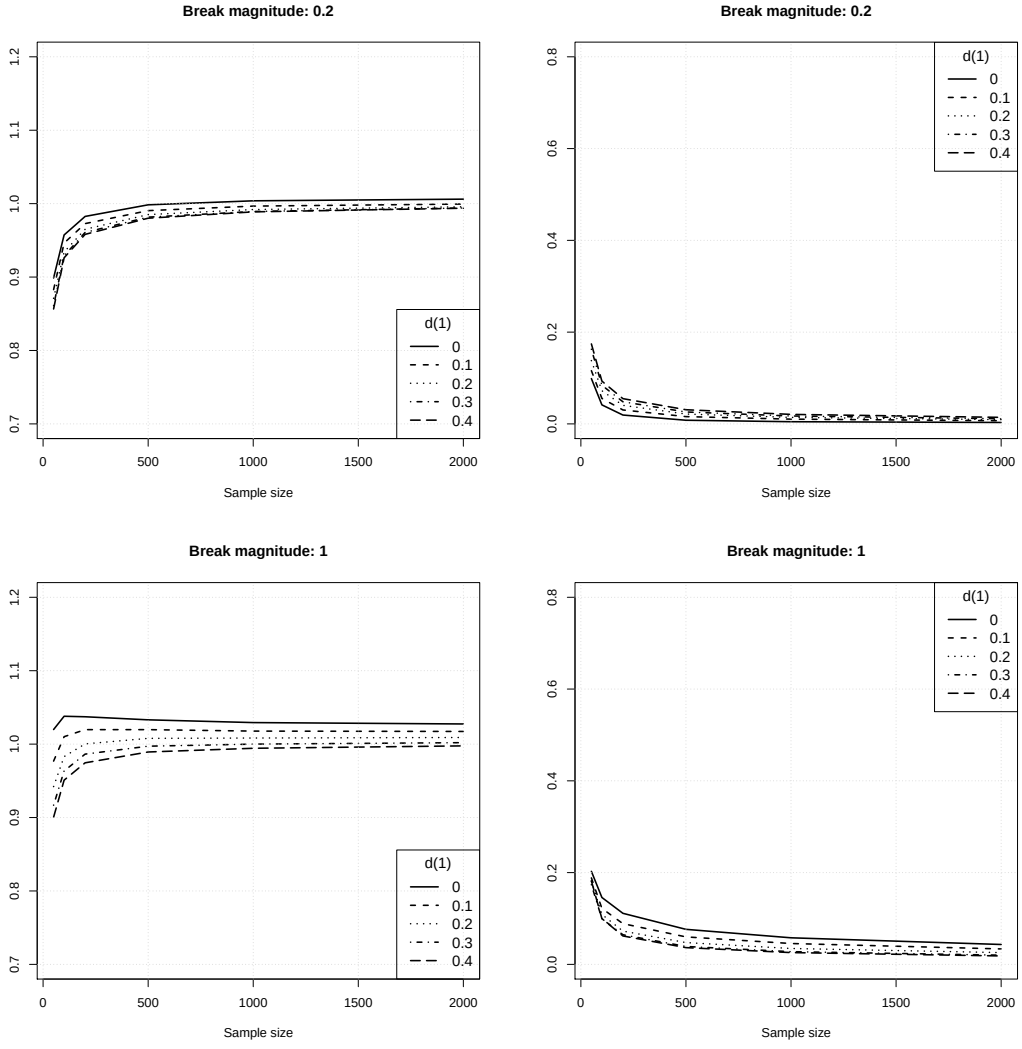

Figure 2: Average of residual variance (left), and variance of difference between true and LAR forecast (right) for $I(d)$ processes with break in mean

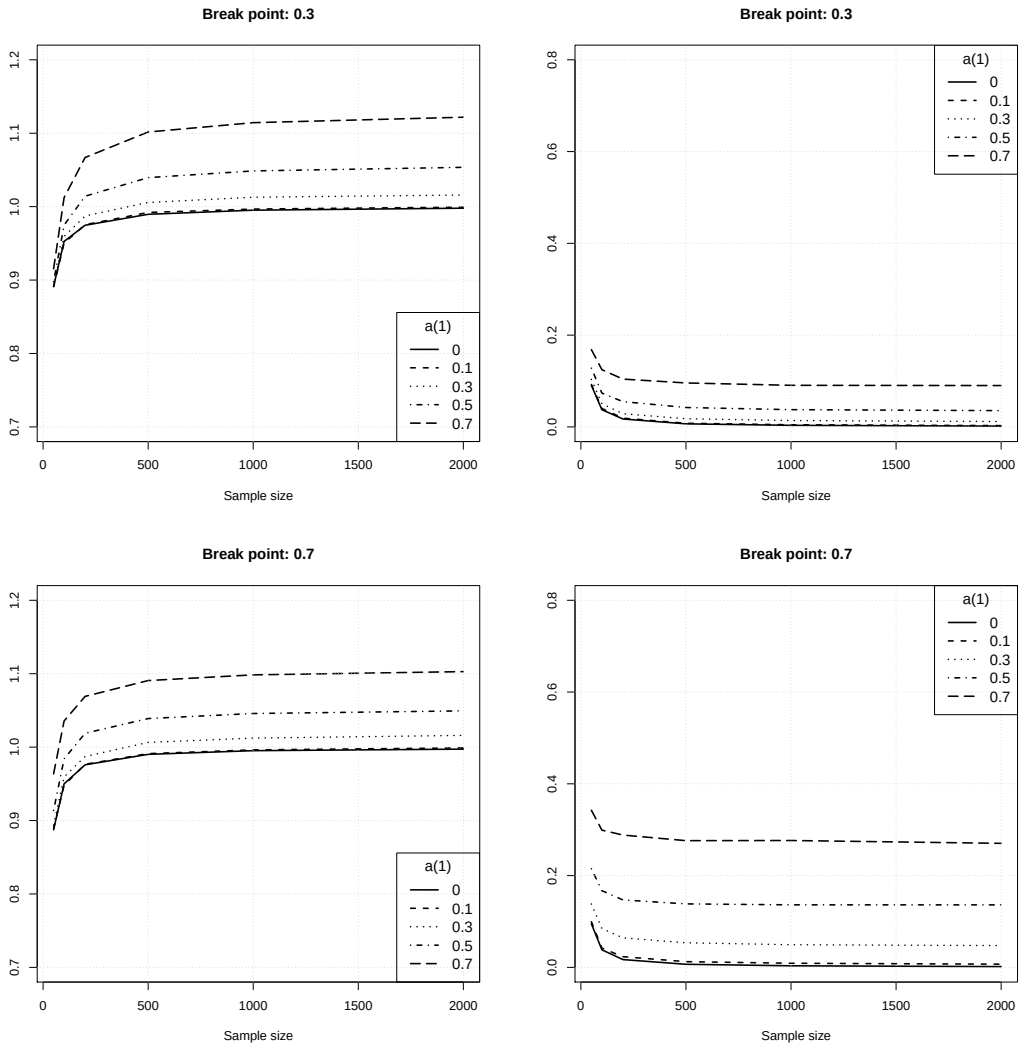

Figure 3: Average of residual variance (left), and variance of difference between true and LAR forecast (right) for AR(1) processes with break in dynamics

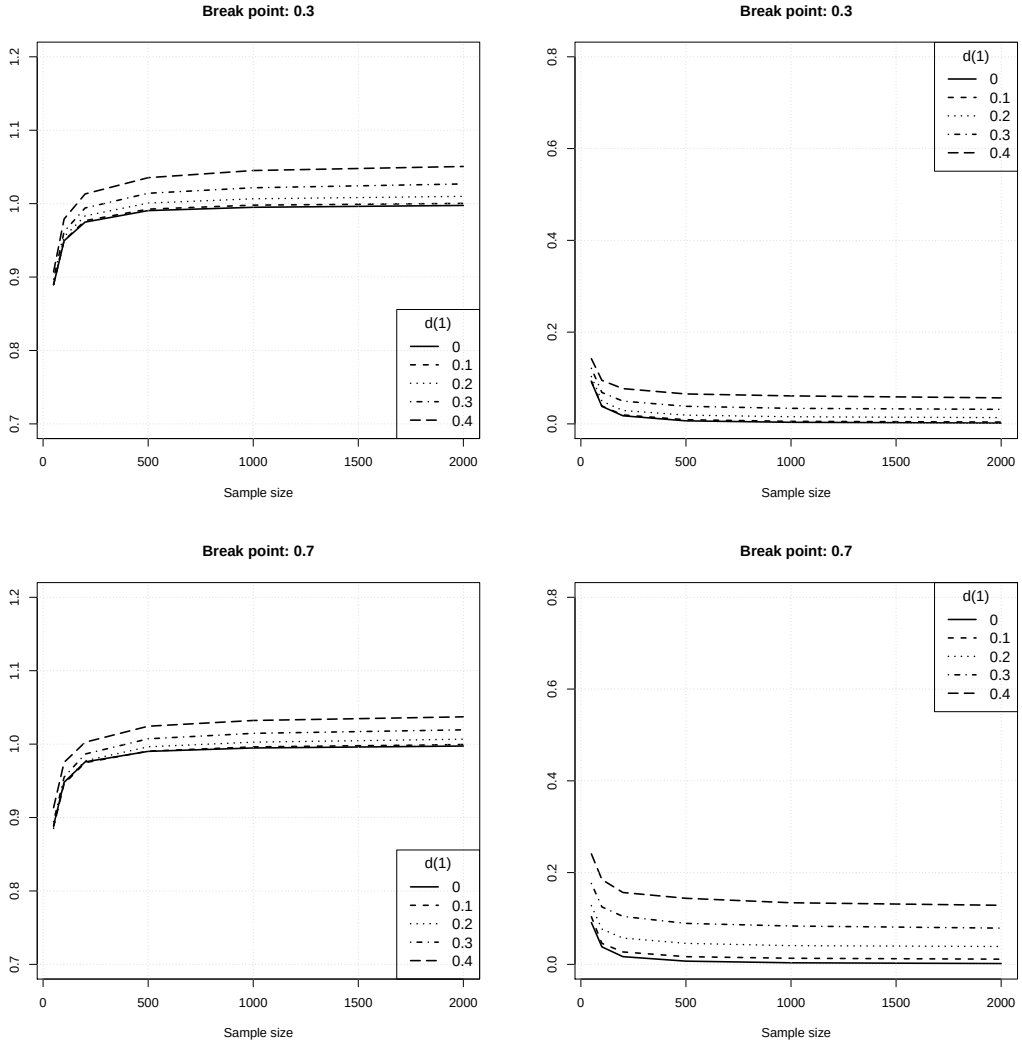

Figure 4: Average of residual variance (left), and variance of difference between true and LAR forecast (right) for $I(d)$ processes with break in dynamics

effect of the size of the autoregressive parameter is rather small.

2. Scenario 2; see Figure 2: For the $I(d)$ case the results are quite similar: For a small break in mean, the in-sample residual variance and variance of $\hat{y}_T(1) - y_T(1)$ are close to what we expect from Proposition 1 and Corollary 1, respectively. With larger breaks in mean, the correspondence is not so close. All in all the graphs very much resemble the ones under Scenario 1. The size of d is of minor importance, and in particular the variance of the differences between the forecasts is close to zero. This confirms the favorable performance of LAR reported by WBH and shows that it extends to larger sample sizes if the parameter break is restricted to the mean.

3. Scenario 3; see Figure 3: For $a_1^{(1)} = a_1^{(2)} = 0$ (no break in dynamics), the in-sample residual variance and the difference between true forecast and LAR forecast converge to 1 and 0, respectively. For $a_1^{(1)} \neq 0$, we know that this is no longer the case (see Proposition 3) which is well illustrated by our experimental evidence. Depending on the size of the AR parameter, the Monte Carlo means and variances converge to different levels. Together with the similar findings to Scenario 4 with fractional integration below, this illustrates once more, that Wang et al. (2013, Theorem 1) is not correct. In particular, when it comes to forecasting (graphs on the right), we observe that a late break fraction ($\tau = 0.7$) induces a stronger bias than an earlier one ($\tau = 0.3$), which is quite intuitive.

4. Scenario 4; see Figure 4: Under long memory, the results from Scenario 3 are essentially reproduced; although, interestingly, the deviations from the theoretical values 1 and 0, respectively, are not as strong as in Figure 3. Still, it is expected that accounting for the break in persistence would improve the forecast performance, see Heinen et al. (2009) for experimental evidence.

6 Concluding remarks

The paper considered the use of long autoregressions for forecasting processes subject to structural change.

A rigorous analysis showed that breaks in the mean, or slowly varying mean functions, are automatically accounted for in the limit. The fitted long autoregression seemingly has a unit root, thus implicity differencing breaks away, while the dynamics is recovered, such that the resulting conditional forecasts converge to the true forecast function. The result holds under infinite-order autoregressive dynamics including long memory. Furthermore, it was shown that long autoregressions do not possess this nice property when the changes are in the dynamics rather than in the mean.

The Monte Carlo experiments confirmed the theoretical findings, illustrating the use and misuse of long autoregressions in practice.

Appendix

A lemma

Before proceeding to the proofs of the propositions, we provide an auxiliary result.

Lemma 1 *Let $\{m_t\}$ satisfy Assumption 4 with $1/4 < \alpha \leq 1$, and $\boldsymbol{m}_{t-h_T} = (m_{t-1}, \ldots, m_{t-h_T})'$. Further, let $x_t = C(L)\varepsilon_t$, $C(L) = \sum_{j=0}^{\infty} c_j L^j$, with $\{\varepsilon_t\}$ from Assumption 1. The sequence $\{c_j\}$ with $c_0 = 1$ is either absolutely summable or $c_j \sim Gj^{d-1}$ for some constant $G > 0$ and $0 < d < 0.5$ as $j \to \infty$. Then, as $T, h_T \to \infty$ and $h_T \leq CT^{\kappa}$ for some $\kappa < 1/3$,*

$$\left\| \frac{1}{T} \sum_{t=h_T+1}^{T} \boldsymbol{m}_{t-h_T} \boldsymbol{m}'_{t-h_T} - \bar{\mu}^2 \boldsymbol{\iota}\boldsymbol{\iota}' \right\| = O\left(\frac{h_T^{1+\alpha}}{T^{\alpha}} \right)$$

and

$$\left\| \frac{1}{T} \sum_{t=h_T+1}^{T} \boldsymbol{m}_{t-h_T} \boldsymbol{x}'_{t-h_T} \right\| = O_p\left(\frac{h_T}{T^{0.5-d}} \right).$$

Proof: For convenience, we subsume the case of absolutely summable $\{c_j\}$ under the case $d = 0$ in what follows.

To prove the first item, it suffices to show that

$$\max_{1 \leq j,k \leq h_T} \left| \frac{1}{T} \sum_{t=h_T+1}^{T} m_{t-j} m_{t-k} - \bar{\mu}^2 \right| = O\left(\frac{h_T^{\alpha}}{T^{\alpha}} \right). \tag{15}$$

Now, for all $1 \leq j, k \leq h_T$,

$$\frac{1}{T} \sum_{t=h_T+1}^{T} \left| m_t^2 - m_{t-j} m_{t-k} \right| \leq C \left(\frac{h_T}{T} \right)^{\alpha} \tag{16}$$

thanks to the piecewise Hölder continuity of ν: while its jump discontinuities may generate nonvanishing differences between m_t^2 and $m_{t-j}m_{t-k}$, there is a finite number thereof and their effect is of order $O\left(\frac{1}{T}\right)$ in the l.h.s. of (16) and thus negligible compared with $\left(\frac{h_T}{T}\right)^{\alpha}$. In a second step we note that

$$\frac{1}{T} \sum_{t=h_T+1}^{T} m_t^2 \to \int_0^1 \nu^2(s)\, \mathrm{d}s = \bar{\mu}^2,$$

where the difference between the average and the integral is of order $O\left(\frac{1}{T^{\alpha}}\right)$ (again, the number of discontinuities is finite and their effect negligible). Summing up, Equation (15) holds and the desired result follows immediately.

To prove the second item, we treat ν as if it were uniformly Hölder continuous of order α, since the finite number of jumps in ν has negligible influence; see above. Also, m_t is bounded on $[0,1]$.

Now, the used matrix norm is bounded by the square root of the product of the maximum row-sum and maximum column-sum norms. The sum of the absolute values of the elements on row k is

$$\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=h_T+1}^{T} m_{t-k}x_{t-j}\right| \leq \sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=h_T+1}^{T} m_{t-j}x_{t-j}\right| + \sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=h_T+1}^{T}\left(m_{t-k}-m_{t-j}\right)x_{t-j}\right|. \tag{17}$$

For the first term on the r.h.s. of (17) we have with the usual convention that $\sum_m^n = 0$ when $m > n$ that

$$\begin{aligned}\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=h_T+1}^{T} m_{t-j}x_{t-j}\right| \;\leq\; & \sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=2}^{T} m_{t-1}x_{t-1}\right| \\ & + \sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=2}^{h_T-j+1} m_{t-1}x_{t-1}\right| + \sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=T-j+2}^{T} m_{t-1}x_{t-1}\right|;\end{aligned}$$

note that the r.h.s. does not depend on k and thus gives an upper bound for the maximum over all rows of $\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=h_T+1}^{T} m_{t-j}x_{t-j}\right|$. Analyzing its behaviour, the variance of $\frac{1}{T}\sum_{t=2}^{T} m_{t-1}x_{t-1}$ is easily checked to be $O\left(T^{2d-1}\right)$ thanks to the boundedness of m_t and the $O\left(h^{2d-1}\right)$ order of the autocovariances of x_t for $0 < d < 0.5$ (and absolute summability for $d = 0$). At the same time,

$$\mathrm{E}\left(\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=2}^{h_T-j+1} m_{t-1}x_{t-1}\right|\right) \leq \frac{\max_{1\leq t\leq T}|m_{t-1}|}{T}\sum_{j=1}^{h_T}\sum_{t=2}^{h_T-j+1}\mathrm{E}\left(|x_{t-1}|\right) \leq C\frac{h_T^2}{T}$$

and analogously

$$\mathrm{E}\left(\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=T-j+2}^{T} m_{t-1}x_{t-1}\right|\right) \leq C\frac{h_T^2}{T}.$$

Thus

$$\max_{1\leq k\leq h_T}\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=h_T+1}^{T} m_{t-j}x_{t-j}\right| = O_p\left(\max\left\{\frac{h_T^2}{T};\frac{h_T}{T^{0.5-d}}\right\}\right) = O_p\left(\frac{h_T}{T^{0.5-d}}\right).$$

For the second term on the r.h.s. of (17), we have that $\max_{1\leq j,k\leq h_T}|m_{t-k}-m_{t-j}|\leq C\left(\frac{h_T}{T}\right)^{\alpha}$ thanks to the Hölder condition on ν, such that

$$\begin{aligned}\operatorname{Var}\left(\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=h_T+1}^{T}\left(m_{t-k}-m_{t-j}\right)x_{t-j}\right|\right) &\leq h_T^2\max_{1\leq k,j\leq h_T}\operatorname{Var}\left(\left|\frac{1}{T}\sum_{t=h_T+1}^{T}\left(m_{t-k}-m_{t-j}\right)x_{t-j}\right|\right)\\ &\leq Ch_T^2\left(\frac{h_T}{T}\right)^{2\alpha}T^{2d-1}.\end{aligned}$$

Now, the maximum over h_T uniformly L_2-bounded variables is of order $O_p\left(\sqrt{h_T}\right)$; by normalizing $\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=h_T+1}^{T}\left(m_{t-k}-m_{t-j}\right)x_{t-j}\right|$ with $h_T\left(\frac{h_T}{T}\right)^{\alpha}T^{d-0.5}$ we may thus conclude that

$$\max_{1\leq k,j\leq h_T}\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=h_T+1}^{T}\left(m_{t-k}-m_{t-j}\right)x_{t-j}\right|=O_p\left(h_T^{1.5}\left(\frac{h_T}{T}\right)^{\alpha}T^{d-0.5}\right)$$

which, for $\alpha>1/4$ and $\kappa<1/3$ is $O_p\left(\frac{h_T}{T^{0.5-d}}\right)$ as can easily be checked (for any $\alpha>1/4$ we have that $\frac{\alpha}{0.5+\alpha}>\frac{1}{3}>\kappa$ as required). Summing up, we have that

$$\max_{1\leq k\leq h_T}\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=h_T+1}^{T}m_{t-k}x_{t-j}\right|=O_p\left(\frac{h_T}{T^{0.5-d}}\right);$$

the same arguments apply for the maximum column-sums norm, such that one has

$$\left\|\frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{m}_{t-h_T}\boldsymbol{x}_{t-h_T}'\right\|\leq\sqrt{\left\|\frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{m}_{t-h_T}\boldsymbol{x}_{t-h_T}'\right\|_1\left\|\frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{m}_{t-h_T}\boldsymbol{x}_{t-h_T}'\right\|_{\infty}}=O_p\left(\frac{h_T}{T^{0.5-d}}\right)$$

as required.

Proof of Proposition 1

Let $\hat{\Sigma}=\frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{y}_{t-h_T}\boldsymbol{y}_{t-h_T}'$ and $\Sigma=\Sigma_{h_T}+\bar{\mu}^2\boldsymbol{\iota}\boldsymbol{\iota}'$, as well as $\hat{\Gamma}=\frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{y}_{t-h_T}y_t$ and $\Gamma=\Gamma_{h_T}+\bar{\mu}^2\boldsymbol{\iota}$. Let again $\boldsymbol{m}_{t-h_T}=(m_{t-1},\ldots,m_{t-h_T})'$.

Note as a preliminary result that, using the Sherman-Morrison formula,

$$\Sigma^{-1}=\left(\Sigma_{h_T}+\bar{\mu}^2\boldsymbol{\iota}\boldsymbol{\iota}'\right)^{-1}=\Sigma_{h_T}^{-1}\left(I-\frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\boldsymbol{\iota}\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\right),$$

implying that

$$\left\|\Sigma^{-1}\right\| \leq \left\|\Sigma_{h_T}^{-1}\right\| \left(1 + \left|\frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\right| \left\|\boldsymbol{\iota}\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\right\|\right).$$

Furthermore $\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota} = Ch_T$ for a suitable $C > 0$ has the same order as $\left\|\boldsymbol{\iota}\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\right\| \leq \|\boldsymbol{\iota}\boldsymbol{\iota}'\| \left\|\Sigma_{h_T}^{-1}\right\| = O(h_T)$, hence

$$\left\|\Sigma^{-1}\right\| = O_p\left(\left\|\Sigma_{h_T}^{-1}\right\|\right) = O_p(1).$$

Turning our attention to the OLS estimator, we have

$$\hat{\boldsymbol{a}}_{h_T} = \hat{\Sigma}^{-1}\hat{\Gamma} = \hat{\Sigma}^{-1}\left(\hat{\Gamma} - \Gamma\right) + \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right)\Gamma + \Sigma^{-1}\Gamma$$

such that

$$\left\|\hat{\boldsymbol{a}}_{h_T} - \Sigma^{-1}\Gamma\right\| \leq \left\|\hat{\Sigma}^{-1}\right\| \left\|\hat{\Gamma} - \Gamma\right\| + \left\|\hat{\Sigma}^{-1} - \Sigma^{-1}\right\| \|\Gamma\|.$$

Obviously, $\|\Gamma\| = O\left(\sqrt{h_T}\right)$. We then need to analyze the remaining norms.

To do so, let us first examine

$$\begin{aligned}
\hat{\Sigma} - \Sigma &= \frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{y}_{t-h_T}\boldsymbol{y}'_{t-h_T} - \Sigma_{h_T} - \bar{\mu}^2\boldsymbol{\iota}\boldsymbol{\iota}' \\
&= \frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{x}_{t-h_T}\boldsymbol{x}'_{t-h_T} - \Sigma_{h_T} + \frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{m}_{t-h_T}\boldsymbol{m}'_{t-h_T} - \bar{\mu}^2\boldsymbol{\iota}\boldsymbol{\iota}' \\
&\quad + \frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{m}_{t-h_T}\boldsymbol{x}'_{t-h_T} + \frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{x}_{t-h_T}\boldsymbol{m}'_{t-h_T}.
\end{aligned}$$

It follows from Lemma 1 that

$$\left\|\frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{m}_{t-h_T}\boldsymbol{x}'_{t-h_T}\right\| = O_p\left(\frac{h_T}{\sqrt{T}}\right).$$

Moreover, given the rate restriction $h_T = O(T^{\kappa})$ for some $\kappa < 1/3$ and the uniformly bounded variance of ε_t^2, we have e.g. from Demetrescu (2009, Lemma 7) that

$$\left\|\frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{x}_{t-h_T}\boldsymbol{x}'_{t-h_T} - \Sigma_{h_T}\right\| = O_p\left(\frac{h_T}{\sqrt{T}}\right),$$

and for the remaining terms we have from Lemma 1 that

$$\left\|\frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{m}_{t-h_T}\boldsymbol{m}'_{t-h_T} - \bar{\mu}^2\boldsymbol{\iota}\boldsymbol{\iota}'\right\| = O\left(\frac{h_T^2}{T}\right)$$

such that

$$\left\|\hat{\Sigma}-\Sigma\right\| = O_p\left(\max\left\{\frac{h_T^2}{T};\frac{h_T}{\sqrt{T}}\right\}\right).$$

As a consequence,

$$\left\|\hat{\Sigma}^{-1}\right\| \leq \left\|\hat{\Sigma}^{-1}-\Sigma^{-1}\right\| + \left\|\Sigma^{-1}\right\| = O_p\left(1\right);$$

moreover, since $\|\Sigma^{-1}\|\left\|\hat{\Sigma}-\Sigma\right\| < 1$, it holds (Lütkepohl, 1996, Section 8.4.1 11(c)) that

$$\left\|\Sigma^{-1}-\hat{\Sigma}^{-1}\right\| \leq \left\|\Sigma^{-1}\right\| \frac{\|\Sigma^{-1}\|\left\|\hat{\Sigma}-\Sigma\right\|}{1-\|\Sigma^{-1}\|\left\|\hat{\Sigma}-\Sigma\right\|}$$

and thus, with $\|\Sigma^{-1}\| < \infty$, it follows that

$$\left\|\hat{\Sigma}^{-1}-\Sigma^{-1}\right\| = O_p\left(\left\|\hat{\Sigma}-\Sigma\right\|\right) = O_p\left(\max\left\{\frac{h_T^2}{T};\frac{h_T}{\sqrt{T}}\right\}\right).$$

Similarly

$$\begin{aligned}
\hat{\Gamma}-\Gamma &= \frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{y}_{t-h_T}y_t - \Gamma_{h_T} - \bar{\mu}^2\boldsymbol{\iota} \\
&= \frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{x}_{t-h_T}x_t - \Gamma_{h_T} + \frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{m}_{t-h_T}m_t - \bar{\mu}^2\boldsymbol{\iota} \\
&\quad + \frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{x}_{t-h_T}m_t + \frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{m}_{t-h_T}x_t.
\end{aligned}$$

We have analogously to the relations above that

$$\left\|\frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{x}_{t-h_T}m_t\right\| = O_p\left(\sqrt{\frac{h_T}{T}}\right) = \left\|\frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{m}_{t-h_T}x_t\right\|,$$

$$\left\|\frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{m}_{t-h_T}m_t - \bar{\mu}^2\boldsymbol{\iota}\right\| = O\left(\frac{h_T^{1.5}}{T}\right);$$

using again the arguments of Demetrescu (2009, Lemma 7), we furthermore obtain

$$\left\|\frac{1}{T}\sum_{t=h_T+1}^{T}\boldsymbol{y}_{t-h_T}y_t - \Gamma_{h_T}\right\| = O_p\left(\sqrt{\frac{h_T}{T}}\right)$$

such that, summing up,

$$\left\|\hat{\Gamma}-\Gamma\right\| = O_p\left(\max\left\{\frac{h_T^{1.5}}{T};\sqrt{\frac{h_T}{T}}\right\}\right).$$

Hence

$$\left\|\hat{\boldsymbol{a}}_{h_T}-\Sigma^{-1}\Gamma\right\| = O_p\left(\max\left\{\frac{h_T^2}{T};\frac{h_T}{\sqrt{T}}\right\}\right)+O_p\left(\max\left\{\frac{h_T^{1.5}}{T};\sqrt{\frac{h_T}{T}}\right\}\cdot\sqrt{h_T}\right).$$

This is $o_p\left(\sqrt{h_T}\right)$ when suitably choosing $\kappa<1/3$.

Focusing now on the "centering" sequence

$$\tilde{\boldsymbol{a}}_{h_T}=\Sigma^{-1}\Gamma=\left(\Sigma_{h_T}+\bar{\mu}^2\boldsymbol{\iota}\boldsymbol{\iota}'\right)^{-1}\left(\Gamma_{h_T}+\bar{\mu}^2\boldsymbol{\iota}\right),$$

use the Sherman-Morrison formula again to obtain that

$$\begin{aligned}
\tilde{\boldsymbol{a}}_{h_T} &= \left(I-\frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\Sigma_{h_T}^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\right)\Sigma_{h_T}^{-1}\Gamma_{h_T}+\bar{\mu}^2\Sigma_{h_T}^{-1}\boldsymbol{\iota}-\frac{1}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\bar{\mu}^2\Sigma_{h_T}^{-1}\boldsymbol{\iota}\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}\\
&= \left(I-\frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\Sigma_{h_T}^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\right)\boldsymbol{a}_{h_T}+\bar{\mu}^2\Sigma_{h_T}^{-1}\boldsymbol{\iota}\left(1-\frac{1}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}\right)\\
&= \boldsymbol{a}_{h_T}-\frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\Sigma_{h_T}^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\boldsymbol{a}_{h_T}+\frac{1}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\bar{\mu}^2\Sigma_{h_T}^{-1}\boldsymbol{\iota}\\
&= \boldsymbol{a}_{h_T}+\frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\Sigma_{h_T}^{-1}\boldsymbol{\iota}\left(1-\boldsymbol{\iota}'\boldsymbol{a}_{h_T}\right)
\end{aligned}$$

as required.

Proof of Corollary 1

By definition we have that

$$\hat{y}_T(1)=\sum_{j=1}^{h_T}\tilde{a}_{j,h_T}y_{T+1-j}+\sum_{j=1}^{h_T}\left(\hat{a}_{j,h_T}-\tilde{a}_{j,h_T}\right)y_{T+1-j}.$$

With $\left\|\boldsymbol{y}_{T+1-h_T}\right\|=\left\|(y_T,\ldots,y_{T+1-h_T})'\right\|=O_p\left(\sqrt{h_T}\right)$ it follows

$$\left|\sum_{j=1}^{h_T}\left(\hat{a}_{j,h_T}-\tilde{a}_{j,h_T}\right)y_{T+1-j}\right|\leq\sqrt{\left\|\hat{\boldsymbol{a}}_{h_T}-\tilde{\boldsymbol{a}}_{h_T}\right\|\left\|\boldsymbol{y}_{T+1-h_T}\right\|}=o_p\left(1\right)$$

such that

$$\hat{y}_T(1) = \sum_{j=1}^{h_T} \tilde{a}_{j,h_T} y_{T+1-j} + o_p(1).$$

Examining the nonnegligible term of $\hat{y}_T(1)$ we further obtain that

$$\sum_{j=1}^{h_T} \tilde{a}_{j,h_T} y_{T+1-j} = \mu_2 \sum_{j=1}^{h_T} \tilde{a}_{j,h_T} + \sum_{j=1}^{h_T} \tilde{a}_{j,h_T} x_{T+1-j}.$$

Since $\sum_{j=1}^{h_T} \tilde{a}_{j,h_T} \to 1$ as $T \to \infty$ by Proposition 1, the correct mean μ_2 at the end of the sample is automatically taken into consideration for the out-of-sample forecast when T is large. With $y_T(1) = \mu_2 + \boldsymbol{x}'_{T+1-h_T} \boldsymbol{a}_{h_T}$ one obtains

$$\sum_{j=1}^{h_T} \tilde{a}_{j,h_T} x_{T+1-j} = \boldsymbol{x}'_{T+1-h_T} \tilde{\boldsymbol{a}}_{h_T} = y_T(1) - \mu_2 + \frac{\bar{\mu}^2}{1 + \bar{\mu}^2 \boldsymbol{\iota}' \Sigma_{h_T}^{-1} \boldsymbol{\iota}} \boldsymbol{x}'_{T+1-h_T} \Sigma_{h_T}^{-1} \boldsymbol{\iota} \left(1 - \boldsymbol{\iota}' \boldsymbol{a}_{h_T}\right).$$

We are thus left with showing that the third summand on the r.h.s. of this equation vanishes as $T \to \infty$. This holds true, since $(1 - \boldsymbol{\iota}' \boldsymbol{a}_{h_T})$ is bounded, see above, and

$$\boldsymbol{x}'_{t-h_T} \Sigma_{h_T}^{-1} \boldsymbol{\iota} = O_p\left(\sqrt{\boldsymbol{\iota}' \Sigma_{h_T}^{-1} \boldsymbol{\iota}}\right)$$

since

$$\begin{aligned} \operatorname{Var}\left(\boldsymbol{x}'_{t-h_T} \Sigma_{h_T}^{-1} \boldsymbol{\iota}\right) &= \boldsymbol{\iota}' \Sigma_{h_T}^{-1} \operatorname{Cov}\left(\boldsymbol{x}_{t-h_T}\right) \Sigma_{h_T}^{-1} \boldsymbol{\iota} \\ &= \boldsymbol{\iota}' \Sigma_{h_T}^{-1} \boldsymbol{\iota}. \end{aligned}$$

At the same time,

$$\frac{\bar{\mu}^2}{1 + \bar{\mu}^2 \boldsymbol{\iota}' \Sigma_{h_T}^{-1} \boldsymbol{\iota}} = O\left(\frac{1}{\boldsymbol{\iota}' \Sigma_{h_T}^{-1} \boldsymbol{\iota}}\right)$$

and the result follows given that $\boldsymbol{\iota}' \Sigma_{h_T}^{-1} \boldsymbol{\iota} \to \infty$. Hence the proof is complete.

Proof of Proposition 2

The steps of the proof are essentially the same as in the proof of Proposition 1 and we use the same notation with $\tilde{\boldsymbol{a}}_{h_T} = \Sigma^{-1}\Gamma$ etc.

Hassler and Kokoszka (2010) show that, if $j^{1-d} b_j \to 0$ with $0 < d < 1$, the Wold coefficients of x_t behave asymptotically as those of the fractional white noise of integration order d. Hence we may build on results derived for fractional white noise for $0 < d < 0.5$, and on the analogous results for absolutely summable coefficients for $d = 0$.

Note that, in the presence of long memory $0 < d < 0.5$, $\gamma_h = \text{Cov}(x_t, x_{t-h}) = O\left(h^{2d-1}\right)$ such that $\|\Sigma_{h_T}\| = O\left(h_T^{2d}\right)$ and $\|\Gamma_{h_T}\| = O\left(h_T^{d}\right)$. Still, $\left\|\Sigma_{h_T}^{-1}\right\| = O(1)$ like in the short-memory case (which is recovered for $d = 0$ of course). Moreover, $\boldsymbol{\iota}\Sigma_{h_T}^{-1}\boldsymbol{\iota}' \geq Ch_T^{1-2d}$ since $\Sigma_{h_T}^{-1}$ is positive definite and its smallest eigenvalue is $O\left(h_T^{-2d}\right)$, implying that $\boldsymbol{\iota}\Sigma_{h_T}^{-1}\boldsymbol{\iota}' \to \infty$ as $T \to \infty$ (and thus $h_T \to \infty$).

It follows as in the proof of Proposition 1 that

$$\left\|\Sigma^{-1}\right\| = O_p\left(\left\|\Sigma_{h_T}^{-1}\right\| \frac{\left\|\boldsymbol{\iota}\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\right\|}{\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\right) = O_p\left(h_T^{2d}\right).$$

Then,

$$\left\|\hat{\boldsymbol{a}}_{h_T} - \Sigma^{-1}\Gamma\right\| \leq \left\|\hat{\Sigma}^{-1}\right\| \left\|\hat{\Gamma} - \Gamma\right\| + \left\|\hat{\Sigma}^{-1} - \Sigma^{-1}\right\| \|\Gamma\|$$

where $\|\Gamma\| \leq \|\Gamma_{h_T}\| + \bar{\mu}^2 \|\boldsymbol{\iota}\| = O\left(\sqrt{h_T}\right)$.

To establish the behavior of the r.h.s. of the above inequality, the same norms as in the proof of Proposition 1 need to be examined.

From Poskitt (2007, Theorem 1) it follows that

$$\left\|\frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{x}_{t-h_T}\boldsymbol{x}'_{t-h_T} - \Sigma_{h_T}\right\| = O_p\left(h_T\left(\frac{\log T}{T}\right)^{0.5-d}\right),$$

since the innovations ε_t satisfy his Assumption 1, and our rate restrictions certainly satisfy his. Using the magnitude orders from Lemma 1, we thus obtain

$$\left\|\hat{\Sigma} - \Sigma\right\| = O_p\left(\max\left\{\frac{h_T^{1+\alpha}}{T^{\alpha}}; h_T\left(\frac{\log T}{T}\right)^{0.5-d}\right\}\right).$$

With both $h_T\left(\frac{\log T}{T}\right)^{0.5-d}$ vanishing and $\frac{h_T^{1+\alpha}}{T^{\alpha}}$ dominated by h_T^{2d} since $\kappa < \frac{\alpha}{\alpha+1.5+2d}$ and $\frac{\alpha}{\alpha+1.5+2d} < \frac{\alpha}{1+\alpha-2d}$, we further have that

$$\left\|\hat{\Sigma}^{-1}\right\| \leq \left\|\hat{\Sigma}^{-1} - \Sigma^{-1}\right\| + \left\|\Sigma^{-1}\right\| = O_p\left(\left\|\Sigma^{-1}\right\|\right) = O_p\left(h_T^{2d}\right).$$

Moving on to the behavior of $\left\|\hat{\Gamma} - \Gamma\right\|$, we exploit the uniform boundedness of the variance of $\frac{1}{T}\sum_{t=h_T+1}^{T} x_{t-j}m_{t-k}$ for $1 \leq j, k \leq h_T$ (implied by boundedness of m_t and weak stationarity of x_t) to conclude that

$$\left\|\frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{x}_{t-h_T}m_t\right\| = O_p\left(\frac{\sqrt{h_T}}{T^{0.5-d}}\right) = \left\|\frac{1}{T}\sum_{t=h_T+1}^{T} \boldsymbol{m}_{t-h_T}x_t\right\|.$$

Lemma 1 further allows us to conclude that

$$\left\| \frac{1}{T} \sum_{t=h_T+1}^{T} \boldsymbol{m}_{t-h_T} m_t - \bar{\mu}^2 \boldsymbol{\iota} \right\| = O\left(\frac{h_T^{0.5+\alpha}}{T^{\alpha}} \right),$$

and, using again Theorem 1 of Poskitt (2007), we have that

$$\left\| \frac{1}{T} \sum_{t=h_T+1}^{T} \boldsymbol{x}_{t-h_T} x_t - \Gamma_{h_T} \right\| = O_p\left(\sqrt{h_T} \left(\frac{\log T}{T} \right)^{0.5-d} \right).$$

Hence

$$\left\| \hat{\Gamma} - \Gamma \right\| = O_p\left(\max\left\{ \frac{h_T^{0.5+\alpha}}{T^{\alpha}}; \sqrt{h_T} \left(\frac{\log T}{T} \right)^{0.5-d} \right\} \right)$$

such that

$$\begin{aligned} \left\| \hat{\boldsymbol{a}}_{p_T} - \Sigma^{-1}\Gamma \right\| &= O_p\left(h_T^{2d} \cdot \max\left\{ \frac{h_T^{1+\alpha}}{T^{\alpha}}; h_T \left(\frac{\log T}{T} \right)^{0.5-d} \right\} \right) \\ &\quad + O_p\left(\max\left\{ \frac{h_T^{1+\alpha}}{T^{\alpha}}; h_T \left(\frac{\log T}{T} \right)^{0.5-d} \right\} \right) \\ &= O_p\left(\max\left\{ \frac{h_T^{2d+\alpha+1}}{T^{\alpha}}; h_T^{2d+1} \left(\frac{\log T}{T} \right)^{0.5-d} \right\} \right). \end{aligned}$$

To obtain the desired convergence rate for $\|\hat{\boldsymbol{a}}_{p_T} - \Sigma^{-1}\Gamma\|$ it suffices to show that

$$\max\left\{ \frac{T^{\kappa(2d+\alpha+1.5)}}{T^{\alpha}}; (\log T)^{0.5-d} \, \frac{T^{\kappa(2d+1.5)}}{T^{0.5-d}} \right\} \to 0,$$

which is indeed implied by our rate restrictions. With $\Sigma^{-1}\Gamma = \left(\Sigma_{h_T} + \bar{\mu}^2 \boldsymbol{\iota}\boldsymbol{\iota}'\right)^{-1} \left(\Gamma_{h_T} + \bar{\mu}^2 \boldsymbol{\iota}\right)$, the first part of the desired result follows using the Sherman-Morrison formula.

To show that Corollary 1 still holds, note that

$$y_T(1) = m_T + \sum_{j=1}^{\infty} a_j x_{T+1-j} = \sum_{j=1}^{\infty} a_j x_{T+1-j} + m_T \tilde{\boldsymbol{a}}'_{h_T} \boldsymbol{\iota} + o_p(1)$$

since the coefficients $\tilde{a}_{j,h_T}$ sum up to 1 whenever $\boldsymbol{\iota}\Sigma_{h_T}^{-1}\boldsymbol{\iota}' \to \infty$ (see the proof of Corollary

1) and indeed $\boldsymbol{\iota}\Sigma_{h_T}^{-1}\boldsymbol{\iota}' \geq Ch_T^{1-2d} \to \infty$ as argued above. At the same time,

$$\begin{aligned}
\hat{y}_T(1) &= \hat{\boldsymbol{a}}_{h_T}'\boldsymbol{m}_{T+1-h_T} + \hat{\boldsymbol{a}}_{h_T}'\boldsymbol{x}_{T+1-h_T} \\
&= \tilde{\boldsymbol{a}}_{h_T}'\boldsymbol{m}_{T+1-h_T} + \tilde{\boldsymbol{a}}_{h_T}'\boldsymbol{x}_{T+1-h_T} + (\hat{\boldsymbol{a}}_{h_T} - \tilde{\boldsymbol{a}}_{h_T})'\boldsymbol{m}_{T+1-h_T} + (\hat{\boldsymbol{a}}_{h_T} - \tilde{\boldsymbol{a}}_{h_T})'\boldsymbol{x}_{T+1-h_T} \\
&= \tilde{\boldsymbol{a}}_{h_T}'\boldsymbol{m}_{T+1-h_T} + \boldsymbol{x}_{T+1-h_T}'\left(\Sigma_{h_T}^{-1}\Gamma_{h_T} + \frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\Sigma_{p_T}^{-1}\boldsymbol{\iota}\left(1 - \boldsymbol{\iota}'\Sigma_{h_T}^{-1}\Gamma_{h_T}\right)\right) + o_p(1)
\end{aligned}$$

since $\|\hat{\boldsymbol{a}}_{h_T} - \tilde{\boldsymbol{a}}_{h_T}\| = o\left(h_T^{-0.5}\right)$ and $\|\boldsymbol{m}_{T+1-h_T}\| = O\left(\sqrt{h_T}\right) = \|\boldsymbol{x}_{T+1-h_T}\|$ thanks to the boundedness of m_t and the uniformly bounded variance of x_t. Then,

$$\begin{aligned}
y_T(1) - \hat{y}_T(1) &= \tilde{\boldsymbol{a}}_{h_T}'\left(\boldsymbol{m}_{T+1-h_T} - m_T\boldsymbol{\iota}\right) \\
&\quad - \left(\left(x_{T+1} - \sum_{j=1}^{\infty} a_j x_{T+1-j}\right) - \left(x_{T+1} - \boldsymbol{x}_{T+1-h_T}'\Sigma_{h_T}^{-1}\Gamma_{h_T}\right)\right) \\
&\quad - \frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}'\Sigma_{h_T}^{-1}\boldsymbol{\iota}}\boldsymbol{x}_{T+1-h_T}'\Sigma_{p_T}^{-1}\boldsymbol{\iota}\left(1 - \boldsymbol{\iota}'\Sigma_{h_T}^{-1}\Gamma_{h_T}\right) + o_p(1).
\end{aligned}$$

The first term is easily shown to vanish, since

$$\left|\tilde{\boldsymbol{a}}_{h_T}'\left(\boldsymbol{m}_{T+1-h_T} - m_T\boldsymbol{\iota}\right)\right| \leq \|\tilde{\boldsymbol{a}}_{h_T}\|\,\|\boldsymbol{m}_{T+1-h_T} - m_T\boldsymbol{\iota}\| \leq C\sqrt{h_T}\left(\frac{h_T}{T}\right)^{\alpha}$$

thanks to the absolute summability of the elements of $\tilde{\boldsymbol{a}}_{h_T}$ and the restrictions on α and κ (cf. the proof of Lemma 1). For the second term, note that $x_{T+1} - \sum_{j=1}^{\infty} a_j x_{T+1-j}$ is the forecast error from a projection of x_t on its infinite past, and $x_{T+1} - \boldsymbol{x}_{T+1-h_T}'\Sigma_{h_T}^{-1}\Gamma_{h_T}$ the forecast error from a projection on its first h_T lags only, and basic Hilbert space arguments show the difference between the two to vanish as $h_T \to \infty$. The third one is shown to vanish as in the proof of Corollary 1, since $\boldsymbol{\iota}\Sigma_{h_T}^{-1}\boldsymbol{\iota}' \to \infty$. Hence, the proof is complete.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics 21*(1), 243–247.

Berenguer-Rico, V. and J. Gonzalo (2014). Summability of stochastic processes: A generalization of integration for non-linear processes. *Journal of Econometrics 178*(2), 331–341.

Berk, K. N. (1974). Consistent autoregressive spectral estimates. *The Annals of Statistics 2*(3), 489–502.

Bhansali, R. (1978). Linear prediction by autoregressive model fitting in the time domain. *The Annals of Statistics 6*(1), 224–231.

Bhattacharya, R. N., V. K. Gupta, and E. Waymire (1983). The Hurst effect under trends. *Journal of Applied Probability 20*, 649–662.

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods*. Springer.

Clements, M. P. and D. F. Hendry (2006). Forecasting with breaks. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 605–657. Elsevier.

Clements, M. P. and D. F. Hendry (2011). Forecasting from misspecified models in the presence of unanticipated location shifts. In M. P. Clements and D. F. Hendry (Eds.), *Oxford Handbook of Economic Forecasting*, pp. 271–313. Oxford University Press.

Demetrescu, M. (2009). Panel unit root testing with nonlinear instruments for infinite-order autoregressive processes. *Journal of Time Series Econometrics 1*(2), article 3.

Diebold, F. X. and A. Inoue (2001). Long memory and regime switching. *Journal of Econometrics 105*(1), 131–159.

Engle, R. F. and A. D. Smith (1999). Stochastic permanent breaks. *Review of Economics and Statistics 81*(4), 553–574.

Giraitis, L., P. Kokoszka, and R. Leipus (2001). Testing for long memory in the presence of a general trend. *Journal of Applied Probability 38*, 1033–1054.

Gonçalves, S. and L. Kilian (2007). Asymptotic and bootstrap inference for ar (infinity) processes with conditional heteroskedasticity. *Econometric Reviews 26*(6), 609–641.

Granger, C. W. J. and N. Hyung (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance 11*(3), 399–421.

Hassler, U. and P. Kokoszka (2010). Impulse responses of fractionally integrated processes with long memory. *Econometric Theory 26*, 1855–1861.

Heinen, F., P. Sibbertsen, and R. Kruse (2009). Forecasting long memory time series under a break in persistence. CREATES Research Paper 2009-53.

Klemeš, V. (1974). The Hurst phenomenon: A puzzle? *Water Resources Research 10*, 675–688.

Lütkepohl, H. (1996). *Handbook of Matrices*. John Wiley & Sons.

Poskitt, D. S. (2007). Autoregressive approximation in nonstandard situations: The fractionally integrated and non-invertible cases. *Annals of the Institute of Statistical Mathematics 59*(4), 697–725.

Poskitt, D. S. (2008). Properties of the sieve bootstrap for fractionally integrated and non-invertible processes. *Journal of Time Series Analysis 29*(2), 224–250.

Ray, B. K. (1993). Modeling long-memory processes for optimal long-range prediction. *Journal of Time Series Analysis 14*(5), 511–525.

Rossi, B. (2013). Advances in forecasting under instability. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, pp. 1203–1324. Elsevier.

Timmermann, A. and H. K. van Dijk (2013). Dynamic econometric modeling and forecasting in the presence of instability. *Journal of Econometrics 177*(2), 131–133.

Wang, C. S.-H., L. Bauwens, and C. Hsiao (2013). Forecasting a long memory process subject to structural breaks. *Journal of Econometrics 177*(2), 171–184.