Econometric Theory, **32**, 2016, 1317–1348.
doi:10.1017/S0266466615000225

(WHEN) DO LONG AUTOREGRESSIONS ACCOUNT FOR NEGLECTED CHANGES IN PARAMETERS?

MATEI DEMETRESCU
Christian-Albrechts-University of Kiel

UWE HASSLER
Goethe University Frankfurt

To construct forecasts for time series exhibiting breaks, the paper examines long autoregressions, where the number of lags is growing with T, and possible breaks are simply ignored. The paper shows that the OLS estimators are still elementwise consistent for the true autoregressive coefficients when neglecting a break in mean, but the sum of the estimators converges to unity. Thanks to this unit-root like behavior of the fitted model, the resulting conditional forecasts are consistent for the true values. As long as the dynamic structure is invariant, the robustness property of the forecasts holds a) under data-dependent lag length selection, b) for a piecewise smoothly varying mean function, and c) under general autoregressive dynamics of possibly infinite order including stationary long memory. Under breaks in the dynamic structure, however, estimators are asymptotically biased, and the forecasts from long autoregressions are biased themselves even in the limit.

1. INTRODUCTION

The dynamic modeling and forecasting of economic and financial time series under breaks in parameters are topics 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). In particular, Phillips (1996) proposes a forecasting framework that addresses both the issue of model selection and "weeding out" data before a parameter change provided the latter is sufficiently important (Phillips, 1996, p. 782). The present paper takes a different route and

The authors would like to thank three anonymous referees as well as Guido Kuersteiner (the deputy editor), Pentti Saikkonen (the co-editor), Andreas Pick, Nazarii Salish and Rob Taylor for very helpful comments that have significantly improved the quality of the paper. Financial support of the German Research Foundation (DFG) through the grants DE 1617/2-1 and HA 3306/5-1 is gratefully acknowledged. The second author further acknowledges financial support by a grant of the VolkswagenStiftung (Opus Magnum). Address correspondence to Matei Demetrescu, Institute for Statistics and Econometrics, Christian-Albrechts-University of Kiel, Olshausenstr. 40-60, D-24118 Kiel, Germany; e-mail: mdeme@stat-econ.uni-kiel.de.

investigates the behavior 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 apparently been first discussed by Durbin (1960) for estimating ARMA models in a two-stage procedure where the unobserved shocks are proxied by first-stage LAR residuals in a feasible ARMA regression. For linear processes with mild summability conditions for the Wold coefficients, Berk (1974) derives the asymptotics of LAR-based spectral density estimators, Bhansali (1978) addresses forecasting, and Gonçalves and Kilian (2007) study bootstrap-based inference. Poskitt (2007, 2008) extends the analysis to long memory and noninvertible processes. Finally, Wang, Bauwens, and Hsiao (2013) (WBH) open the floor for a discussion of LAR under breaks.

Our paper contains two contributions with respect to LAR when the true model is autoregressive and subject to structural changes, where the date or size of the break is not estimated but simply ignored. First, we address the situation of an ignored mean shift under constant dynamics. If the process is autoregressive of finite order, the LAR forecast converges to the conditional mean as true forecast function (Proposition 1 and Corollary 1). In fact, this result continues to hold if the lag length is not chosen deterministically but data-driven according to an information criterion (Proposition 2). Further, Proposition 1 extends to more general conditions: we allow for a mean function with several breaks under AR dynamics of infinite order that may even display long memory. As long as the dynamic structure is invariant over time, the LAR forecast is unbiased for the conditional mean asymptotically (Proposition 3). Second, we turn to the case of breaks in the autoregressive parameters. Here it turns out that the LAR forecasts are conditionally biased in the limit and miss the true forecast function (Remark 3). We present experimental evidence with growing sample sizes illustrating our theoretical results.

The remainder of the paper is structured as follows. In the next section, we specify the details of the model under breaks in parameters and discuss the theoretical statements by WBH. 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, $\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}'\boldsymbol{a}}$, as well as the corresponding induced matrix norm, $\|A\| = \max_{\|\boldsymbol{a}\|=1} \|A\boldsymbol{a}\|$, $\lfloor x \rfloor$ denotes the integer part of a positive number x, probabilistic Landau symbols $o_p(\cdot)$ and $O_p(\cdot)$ have their usual meaning, $\sim$ denotes asymptotic equivalence of two sequences, C stands for a generic constant whose value may change from

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

occurrence to occurrence, and $\xrightarrow{p}$ stands for convergence in probability as the sample size T goes to infinity.

2. MODEL

We work with T observations of a univariate process $\{y_t\}$ 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, allowing for changes in the autoregressive dynamic at the same time:

$$y_t = \begin{cases} \mu_1 + x_t^{(1)}, t = 1, \ldots, T_1 = \lfloor \tau \, T \rfloor \\ \mu_2 + x_t^{(2)}, \quad t = T_1 + 1, , \ldots, T \end{cases}, \quad x_t^{(r)} = A_r^{-1}(L)\varepsilon_t = \sum_{j=0}^{\infty} c_j^{(r)} \varepsilon_{t-j}. \tag{2}$$

To distinguish the two regimes, we use superscripts or subscripts $r \in \{1, 2\}$ with break fraction $\tau \in (0, 1)$. For $\left\{x_t^{(r)}\right\}$ to be weakly stationary, we maintain square summability of the moving average expansion of the inverted autoregressive polynomials $A_r(L)$ defined in terms of the usual lag operator L: $\sum_{j=0}^{\infty} \left(c_j^{(r)}\right)^2 < \infty$. With the autoregressive polynomials A_r, one obtains from (2) for the respective regimes

$$A_r(L) y_t = A_r(1)\mu_r + \varepsilon_t, \quad A_r(L) = 1 - \sum_{j=1}^{\infty} a_j^{(r)} L^j, \ \sum_{j=1}^{\infty} \left(a_j^{(r)}\right)^2 < \infty.$$

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 $\{\varepsilon_t\}_{t \in \mathbb{Z}}$ is $iid\left(0, \sigma^2\right)$ with finite 4th-order moments.

It is noteworthy that $\{y_t\}$ from (2) has no stationary autoregressive representation either in the case of a mean shift or in the presence of a break in the dynamics. Filtering the process with any $A_*(L)$ results in

$$A_*(L) y_t = \begin{cases} A_*(1)\mu_1 + A_*(L) A_1^{-1}(L)\varepsilon_t, \, t \leq T_1 \\ A_*(1)\mu_2 + A_*(L) A_2^{-1}(L)\varepsilon_t, \, t > T_1 \end{cases}. \tag{3}$$

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

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

When it comes to forecasting, we follow the proposal by WBH and do not model the date or type of break. We simply ignore the possibility of breaks and run instead a LAR of y_t estimated by OLS: $y_t = \hat{m} + \sum_{j=1}^{h_T} \hat{a}_{j,h_T} y_{t-j} + \hat{\varepsilon}_t$, $t = h_T + 1, \ldots, T$. With $\hat{\boldsymbol{a}}_{h_T}$ denoting the vector of OLS estimators and $\boldsymbol{y}_{t-h_T} = \big(y_{t-1}, \ldots, y_{t-h_T}\big)'$, we write more compactly

$$y_t = \hat{m} + \hat{\boldsymbol{a}}'_{h_T} \boldsymbol{y}_{t-h_T} + \hat{\varepsilon}_t \,. \tag{4}$$

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 one step ahead forecast function implied by the fitted model is then

$$\hat{y}_T(1) = \hat{m} + \hat{\boldsymbol{a}}'_{h_T} \boldsymbol{y}_{T+1-h_T} = \hat{m} + \sum_{j=1}^{h_T} \hat{a}_{j,h_T} y_{T+1-j}, \tag{5}$$

which is the natural choice given the adopted framework. At the same time, the true forecast function is the conditional mean: $y_T(1) = \mathrm{E}(y_{T+1} | y_T, y_{T-1}, \ldots)$. Note that $y_T(1)$ will depend in general on all the parameters contained in μ_r and $A_r(L)$, $r \in \{1, 2\}$, of the model, which is not known in practice.

The use of long autoregressions like in (4) has been advocated by WBH assuming fractionally integrated noise in (2), i.e.

$$A_r(L) = (1-L)^{d_r} = 1 - \sum_{j=1}^{\infty} a_{j,d}^{(r)} L^j, \quad |d_r| < \frac{1}{2}, \tag{6}$$

with $a_{j,d}^{(r)} = -\pi_{j,d}^{(r)}$, where $\pi_{0,d}^{(r)} = 1$ and $\pi_{j,d}^{(r)} = \frac{j-1-d_r}{j} \pi_{j-1,d}^{(r)}$ with $\pi_{j,d}^{(r)} \sim \frac{j^{-d_r-1}}{\Gamma(-d_r)}$, $j \to \infty$. For simplicity, we assume for now that $\mu_1 = \mu_2 = 0$. Wang et al. (2013, Lemma 1) state that $\{y_t\}$ from (2) with (6) has a representation as a fractionally integrated process without break,

$$(1-L)^{d^*} y_t = \varepsilon_t, \quad d^* = \lambda d_1 + (1-\lambda) d_2, \quad \lambda \in [0, 1] \,, \tag{7}$$

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

$$(1-L)^{d^*} y_t = y_t - \sum_{j=1}^{\infty} a_j^{(*)} y_{t-j} = \varepsilon_t, \tag{8}$$

with the autoregressive coefficients $a_j^{(*)}$ taken from the expansion of $(1-L)^{d^*}$. This statement is not correct as can be seen from (3). In fact, differencing y_t from (2) under (6) results under $\mu_1 = \mu_2 = 0$ in

$$(1-L)^{d^*} y_t = \begin{cases} (1-L)^{d^*-d_1} \varepsilon_t \sim I(d_1 - d^*), & t \le T_1 \\ (1-L)^{d^*-d_2} \varepsilon_t \sim I(d_2 - d^*), & t > T_1 \end{cases} . \tag{9}$$

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

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

Wang et al. (2013, Thm. 1) address the behaviour of OLS estimators from (4). In the model with break in the autoregressive parameters, the OLS estimators cannot converge to true AR parameters simply because an AR representation as in (8) does not exist. Hence, the question raised in the title of this paper comes in naturally: in what situation do LAR actually account for neglected 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 contains the AR(∞) case and long memory. Further, we first assume the number of endogenous regressors h_T to grow deterministically at a controlled rate, while a selection with an information criterion is addressed subsequently.

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,

$$y_t = m_t + x_t\,, \tag{10}$$

where the deterministic mean function m_t exhibits a jump. To simplify matters, we assume a demeaned structural break m_t,

$$m_t = \begin{cases} \mu_1 = -(1-\tau)(m_2 - m_1)\,, & t \le \tau T \\ \mu_2 = \tau\,(m_2 - m_1)\,, & t > \tau T \end{cases}, \tag{11}$$

and consequently, we do not have to allow for an intercept in the long autoregression (4) 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 $A(L)\,x_t = x_t - \sum_{j=1}^{p} a_j x_{t-j} = \varepsilon_t$, $t \in \mathbb{Z}$, where $\{\varepsilon_t\}$ is from Assumption 1, and $A(z)$ has all roots outside the unit circle.

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 (10), the extended model

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

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.

Let $\Sigma_{h_T} = \mathrm{Cov}\left(\boldsymbol{x}_{t-h_T}\right)$ denote the h_Tth-order covariance matrix of $\{x_t\}$ where $\boldsymbol{x}_{t-h_T} = \left(x_{t-1}, \ldots, x_{t-h_T}\right)'$; let also $\Gamma_{h_T} = \mathrm{E}\left(\boldsymbol{x}_{t-h_T} x_t\right)$. 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

$$\left\| \Sigma_{h_T} \right\| = O(1) \quad \text{and} \quad \left\| \Sigma_{h_T}^{-1} \right\| = O(1); \tag{12}$$

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 $\{y_t\}$ satisfy (10) with (11) and Assumption 2. Denote the vector of true parameters in $\mathbb{R}^{h_T}$ by $\boldsymbol{a}_{h_T} = \left(a_1, \ldots, a_p, 0, \ldots, 0\right)'$ and consider the vector of OLS estimators $\hat{\boldsymbol{a}}_{h_T}$ from (4). Further, let*

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

where $\boldsymbol{\iota}_{h_T}$ 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 $\kappa \in \left(0, \frac{1}{4}\right]$, it holds as $T \to \infty$ that

$$\left\| \hat{\boldsymbol{a}}_{h_T} - \tilde{\boldsymbol{a}}_{h_T} \right\| = o_p\left(h_T^{-1/2}\right).$$

Proof. See the Appendix. ■

The rate restriction $\kappa \leq 1/4$ is stricter than the one given by Berk (1974) for obtaining consistent spectral density estimators and is due to the presence of a break in the mean function not accounted for in the LAR. The intuition behind the rate reduction is that elementwise negligible terms involving the ignored break cumulate over $\hat{a}_{j,h_T}$ such that h_T must be reduced in order to maintain the desired first-order limiting behavior.

The sequence $\tilde{\boldsymbol{a}}_{h_T}$ forms a triangular array, and $\tilde{a}_{j,h_T}$ changes for fixed j with the sample size. The closeness of $\tilde{\boldsymbol{a}}_{h_T}$ and $\boldsymbol{a}_{h_T}$ depends on the magnitude and the timing of the jump through $\bar{\mu}^2$, where the effect of the break point is symmetric about $1/2$. E.g., for the special case where $\{x_t\}$ is white noise we have for large h_T

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

where the larger the ratio $\bar{\mu}/\sigma$, the better the approximation. The particular case of white noise with change in mean nicely illustrates the first-order limiting properties of $\hat{\boldsymbol{a}}_{h_T}$ discussed in the following two remarks.

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

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 a) the row sums of $\Sigma_{h_T}^{-1}$ are bounded and b) $\Sigma_{h_T}^{-1}$ has eigenvalues bounded away from zero such that $\boldsymbol{\iota}_{h_T}' \Sigma_{h_T}^{-1} \boldsymbol{\iota}_{h_T} \sim C h_T$. In fact, L_2 convergence of the vector of OLS estimates to the true parameters is also given, but at a lower rate: $\left\| \hat{\boldsymbol{a}}_{h_T} - \boldsymbol{a}_{h_T} \right\|$ is of order $O_p\left(h_T^{-1/2}\right)$ but not $o_p\left(h_T^{-1/2}\right)$. This is because $\left\| \hat{\boldsymbol{a}}_{h_T} - \boldsymbol{a}_{h_T} \right\| \leq \left\| \hat{\boldsymbol{a}}_{h_T} - \tilde{\boldsymbol{a}}_{h_T} \right\| + \left\| \tilde{\boldsymbol{a}}_{h_T} - \boldsymbol{a}_{h_T} \right\|$ and $\left\| \hat{\boldsymbol{a}}_{h_T} - \tilde{\boldsymbol{a}}_{h_T} \right\| \leq \left\| \hat{\boldsymbol{a}}_{h_T} - \boldsymbol{a}_{h_T} \right\| + \left\| \tilde{\boldsymbol{a}}_{h_T} - \boldsymbol{a}_{h_T} \right\|$ where

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

Some algebra shows the bias term $\tilde{\boldsymbol{a}}_{h_T} - \boldsymbol{a}_{h_T}$ to behave as $\frac{1 - \boldsymbol{\iota}_{h_T}' \boldsymbol{a}_{h_T}}{h_T} \boldsymbol{\iota}_{h_T}$ for large h_T.

Table 1 illustrates the convergence behavior of the LAR OLS estimators in a preliminary Monte Carlo experiment. (The design is as follows: for each sample size $T \in \{50, 100, 200, 500, 1{,}000, 2{,}000\}$, we generate 25,000 replications of standard normal iid series and add a centered mean component with a break at $\tau = 1/2$, i.e. $\mu_1 = -1.25$ and $\mu_2 = 1.25$; the fitted autoregression is of order $\lfloor 4(T/100)^{0.25} \rfloor$ and does not include an intercept.) We note that, for very small T, the average estimates may appear to come from an I(d) process. But already for

TABLE 1. Mean and standard deviation (in parentheses) of LAR OLS estimators for a white noise process with break in mean

T	$\bar{\hat{a}}_1$	$\bar{\hat{a}}_2$	$\bar{\hat{a}}_3$	$\bar{\hat{a}}_4$	$\bar{\hat{a}}_5$	$\bar{\hat{a}}_6$	$\bar{\hat{a}}_7$	$\bar{\hat{a}}_8$	$\overline{\sum \hat{a}_j}$
50	0.334	0.230	0.215	–	–	–	–	–	0.778
	(0.130)	(0.153)	(0.127)						(0.065)
100	0.266	0.211	0.192	0.167	–	–	–	–	0.836
	(0.0970)	(0.095)	(0.102)	(0.092)					(0.037)
200	0.242	0.214	0.203	0.191	–	–	–	–	0.849
	(0.068)	(0.064)	(0.072)	(0.064)					(0.024)
500	0.192	0.179	0.175	0.167	0.167	–	–	–	0.881
	(0.045)	(0.043)	(0.048)	(0.045)	(0.042)				(0.012)
1,000	0.142	0.136	0.133	0.128	0.126	0.124	0.124	–	0.913
	(0.032)	(0.031)	(0.031)	(0.034)	(0.032)	(0.031)	(0.030)		(0.007)
2,000	0.123	0.120	0.118	0.115	0.114	0.112	0.112	0.111	0.924
	(0.023)	(0.022)	(0.022)	(0.021)	(0.023)	(0.023)	(0.022)	(0.021)	(0.004)

Note: The innovations are independent standard normal. The break is located in the middle of the sample and is of size $\mu_2 - \mu_1 = 2.5$. The figures are computed as mean and standard deviation over 25,000 Monte Carlo replications. The model order is deterministic, $\lfloor 4(T/100)^{0.25} \rfloor$.

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

$T = 500$, they are fairly close to $1/h_T$ as predicted by Remark 1, and the difference is reduced further with larger T. In fact, for any of the studied sample sizes, the first autoregressive estimator $\hat{a}_{1,h_T}$ is practically equal to $1/h_T$, and only for $j > 2$ does one observe some downward bias in $\hat{a}_{j,h_T}$. Furthermore, it can be seen that all autoregressive estimates tend to become smaller on average as T increases, but apparently more slowly than the respective standard deviations, illustrating the difference between the true parameter values and the centering sequence $\tilde{\boldsymbol{a}}_{h_T}$. The sum of the autoregressive estimators also gets closer to unity as T grows.

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 of the story. The remark does not explain why the neglected mean shift would not affect the forecasts, which should after all be centered at the post-break mean. The following remark sheds light on this issue.

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

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

since $1 - \boldsymbol{\iota}_{h_T}' \boldsymbol{a}_{h_T}$ is bounded thanks to the stability of $\{x_t\}$. In other words, the fitted LAR seemingly have 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. This convergence is nicely illustrated in Table 1, see the last column.

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

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

We have the following result.

COROLLARY 1. *Under the assumptions of Proposition 1, it holds that* $\hat{y}_T(1) = y_T(1) + o_p(1)$ *as* $T \to \infty$.

Proof. See the Appendix. ■

The date and size of an eventual break are unknown in practice and have to be estimated. Such estimates may be quite imprecise leading to deteriorated forecasts. Pesaran and Timmermann (2007) showed that the use of pre-break data may improve forecasts without specifying the break for the forecast exercise. Corollary 1 shows indeed that a long autoregression, ignoring possible breaks results in consistent forecasts (as long as the break is restricted to the mean function and does not affect the dynamics, see Remark 3 below).

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

Also, changes in the mean may also occur at the beginning or the end of the sample. In terms of parameter estimation, they are asymptotically negligible, yet a break occurring at the end of the sample is forecast-relevant. But if there is another break in the data, one for which the asymptotics of Proposition 1 is relevant, the pseudo unit-root behavior of the fitted LAR would also take care of the break at the end of the sample.

Finally, it would be interesting to compare the procedure studied here with some more generic robust forecasting approaches. Robustness to breaks (assuming that one is not modeling and timing them explicitly) can be achieved by resorting to adaptively downweighting data such that only relevant periods serve for setting up the desired forecast. E.g., Giraitis, Kapetanios, and Price (2013) work with flexible weighted averages of past values and show that picking the corresponding bandwidth parameter in real time via cross-validation allows for robustification against breaks: essentially, the procedure uses the whole past before the break and gradually increases the weights attached to postbreak data until prebreak data do not matter. In comparison, with the fitted autoregressive coefficients adding up to unity in the limit, $\hat{y}_T(1)$ is ultimately a weighted average as well, in fact it is a *local* weighted average since $p_T = o(T)$, though not an adaptive one as the one studied in Giraitis et al. (2013).

Summing up, Proposition 1 offers practitioners a safety net when working in an environment where changes in the mean are not excluded a priori. In applied work, however, it is customary to choose the model order using data-driven methods like the Akaike information criterion. Therefore, we address its behavior next.

PROPOSITION 2. *Let $AIC(\ell)$ denote Akaike's information criterion computed from a fitted OLS autoregression of order ℓ,*

$$y_t = \sum_{j=1}^{\ell} \hat{a}_{j,\ell} y_{t-j} + \hat{\varepsilon}_t, \quad 1 \le \ell \le h_T.$$

Under the assumptions of Proposition 1, it holds for $\mu_1 \neq \mu_2$ that $\underset{1 \le \ell \le h_T}{\arg\min}\, AIC(\ell) \xrightarrow{p} \infty$ as $T \to \infty$.

Proof. See the Appendix. ■

From Proposition 2, we learn that the model order selected by Akaike's information criterion satisfies the theoretical rate restrictions on the maximal lag length h_T required by Proposition 1. Hence, the statements of Proposition 1 and Corollary 1 continue to hold when the lag order is determined by AIC.

The interaction of model selection and estimation is illustrated in Table 2, which replicates Table 1 with the added twist of selecting the model order via AIC (the maximal model order is chosen as $\lfloor 4(T/100)^{0.25} \rfloor$). The results are virtually the same for the two tables, which indicate that the maximal model order

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

TABLE 2. Mean and standard deviation (in parentheses) of LAR OLS estimators for a white noise process with break in mean and model selection via AIC

T	$\bar{\hat{a}}_1$	$\bar{\hat{a}}_2$	$\bar{\hat{a}}_3$	$\bar{\hat{a}}_4$	$\bar{\hat{a}}_5$	$\bar{\hat{a}}_6$	$\bar{\hat{a}}_7$	$\bar{\hat{a}}_8$	$\overline{\sum \hat{a}_j}$
50	0.359	0.242	0.156	–	–	–	–	–	0.756
	(0.143)	(0.166)	(0.164)						(0.084)
100	0.278	0.223	0.197	0.129	–	–	–	–	0.827
	(0.099)	(0.105)	(0.111)	(0.122)					(0.043)
200	0.244	0.216	0.205	0.183	–	–	–	–	0.848
	(0.069)	(0.067)	(0.073)	(0.080)					(0.025)
500	0.192	0.180	0.175	0.168	0.166	–	–	–	0.881
	(0.044)	(0.042)	(0.048)	(0.044)	(0.043)				(0.013)
1,000	0.143	0.136	0.133	0.128	0.126	0.124	0.124	–	0.913
	(0.032)	(0.032)	(0.030)	(0.034)	(0.032)	(0.031)	(0.031)		(0.007)
2,000	0.123	0.119	0.117	0.115	0.114	0.112	0.112	0.111	0.924
	(0.023)	(0.022)	(0.022)	(0.021)	(0.024)	(0.023)	(0.022)	(0.022)	(0.004)

Note: The maximal model order given by $\lfloor 4(T/100)^{0.25} \rfloor$. Coefficients not selected are treated as zeros in computing the averages. For further details see Table 1.

is chosen in most of the cases, and that information criteria are a reliable tool for practitioners in this framework.

3.2. Extensions

We now extend the model (10) 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 (11) 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 much more general conditions, and the robustness property from Corollary 1 carries over.

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 $j \to \infty$.

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.

Now, we turn to the mean process. There is in fact no a priori reason to assume just one single break in (11); we may allow, more generally, for several such

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

discontinuities. Moreover, $\{m_t\}$ does not have to be constant between two breaks; we only require continuity between two break times, more precisely only Hölder continuity of some order α. For a function $\nu(\cdot)$ on (a, b), we hence assume

$$\sup_{a<s<t<b} \frac{|\nu(t)-\nu(s)|}{|t-s|^{\alpha}} < \infty \quad \text{for suitable } \alpha \in (0,1].$$

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

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 $\int \nu(s)\,\mathrm{d}s = 0$ implies that the process is demeaned. Hence, we consider again a LAR without an intercept without loss of generality.

PROPOSITION 3. *Consider $\{y_t\}$ from (10) 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 $0 < \kappa < \min\left\{\frac{\alpha}{2+\alpha+4d}; \frac{1-2d}{4+8d}\right\}$, it follows that*

$$\left\|\hat{\boldsymbol{a}}_{h_T} - \tilde{\boldsymbol{a}}_{h_T}\right\| = o_p\left(h_T^{-1/2}\right)$$

as $T \to \infty$, where with $\bar{\mu}^2$ from Assumption 4 one defines

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

Proof. See the Appendix. ■

In the AR(p) case, $\Sigma_{h_T}^{-1}\Gamma_{h_T}$ gives the true autoregressive coefficients for any $h_T \geq p$. This does not hold anymore in the AR(∞) case: rather, $\Sigma_{h_T}^{-1}\Gamma_{h_T}$ gives 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 choice of κ is more limited than in Proposition 1. On the one hand, the presence of long memory imposes $\kappa < \frac{1-2d}{4+8d}$. Analogously, this is stricter than the rate derived for the case without breaks (Poskitt, 2007). The additional restriction $\kappa < \frac{\alpha}{2+\alpha+4d}$ is influenced by the smoothness (or rather roughness) condition on the mean function ν; it is not binding, for instance, when ν satisfies a Lipschitz condition, i.e., when $\alpha = 1$. This 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 $\frac{\alpha_{\min}}{3+4d}$ for some $\alpha_{\min} > 1/4$ that one is prepared to accept. The "worst-case" scenario would be $\kappa < 1/20$ 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 $\frac{1-2d}{4+8d}$ that is binding: for $d > 1/3$, $\frac{1-2d}{4+8d} < \frac{1}{20}$. A logarithmic rate for h_T satisfies both.

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

In spite of the restrictions on the maximal model order h_T caused by the generality of the DGP considered in Proposition 3, Corollary 1 continues to hold. The true one-step ahead conditional forecast at the end of the sample is now given by

$$y_T(1) = \mathrm{E}(y_{T+1}|y_T, y_{T-1}, \ldots) = m_T + \sum_{j=1}^{\infty} a_j x_{T+1-j}.$$

We then have the following

COROLLARY 2. *Under the assumptions of Proposition 3, it holds that* $\hat{y}_T(1) = y_T(1) + o_p(1)$ *as* $T \to \infty$.

Proof. See the Appendix. ■

We hold it for obvious that the model order selected with AIC would go to infinity irrespective of the variations in the mean when the stochastic component is AR(∞), so we do not explicitly formulate the result analogous to Proposition 2 here.

4. BREAKS IN THE AUTOREGRESSIVE COEFFICIENTS

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

$$y_t = \begin{cases} x_t^{(1)} = A_1^{-1}(L)\varepsilon_t, & t \le \tau T \\ x_t^{(2)} = A_2^{-1}(L)\varepsilon_t, & t > \tau T \end{cases}, \tag{13}$$

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, see (5), but without intercept. The process from (13) 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 4. *Let* $\{y_t\}$ *be from (13) and* $\left\{x_t^{(r)}\right\}$, $r = 1, 2$, *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*

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

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

where $\Sigma_{h_T,r}$ *denotes the regime-specific covariance matrix of* $\boldsymbol{x}_{t-h_T}$ *for* $r = 1, 2$. *If* $h_T^{-1} + h_T T^{-\kappa} \to 0$ *for some* $\kappa \in (0, 1/4]$, *it holds as* $T \to \infty$ *that* $\left\| \hat{\boldsymbol{a}}_{h_T} - \bar{\boldsymbol{a}}_{h_T} \right\| = o_p\left(h_T^{-1/2}\right)$.

Proof. Analogous to the proof of Proposition 1 and omitted. ■

We no longer have any kind of 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 $I_{h_T} + \frac{1-\tau}{\tau} \Sigma_{h_T,1}^{-1}$ 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 may be of interest to add some intuition to Proposition 4. 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 $\left\{ \varepsilon_t^{(1)} \right\}$ is independent of $\left\{ \varepsilon_t^{(2)} \right\}$. Then for any fixed autoregressive order $\bar{p}$, the limit of the OLS autoregressive estimators for y_t is given by

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

with $\Sigma_{\bar{p},r}$ and $\Gamma_{\bar{p},r}$ as implied by the lag polynomials A_r, $r = 1, 2$. 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 4. 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.

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

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 $T \in \{50, 100, 200, 500, 1{,}000, 2{,}000\}$ 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 25,000 replications for each parameter constellation. The lag length of the LAR is chosen by Akaike's information criterion, AIC, with a maximum order given by $\lfloor 4(T/100)^{0.25} \rfloor$. We report a) the in-sample residual variance averaged over the 25,000 replications, and b) the variance over 25,000 replications of the difference between the fitted forecast function $\hat{y}_T(1)$ and the true forecast function $y_T(1)$. Both are reported here, 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 as follows for the four scenarios.

1. For the AR(1) process with a break in the mean, we simulate with an autoregressive parameter $a_1 \in \{0, 0.3, 0.5, 0.7, 0.95\}$. The break fraction is taken to be $\tau = 1/2$, and the magnitude of the break is either small, $\mu_2 - \mu_1 = 0.2$, or large, $\mu_2 - \mu_1 = 2.5$; the mean function is centered according to (11).
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 dynamics breaks from $a_1^{(1)} \in \{0, 0.3, 0.5, 0.7, 0.95\}$ for the prebreak sample to independent white noise ($a_1^{(2)} = 0$) for the postbreak period.
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 ($\mu_2 - \mu_1 = 2.5$), the

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

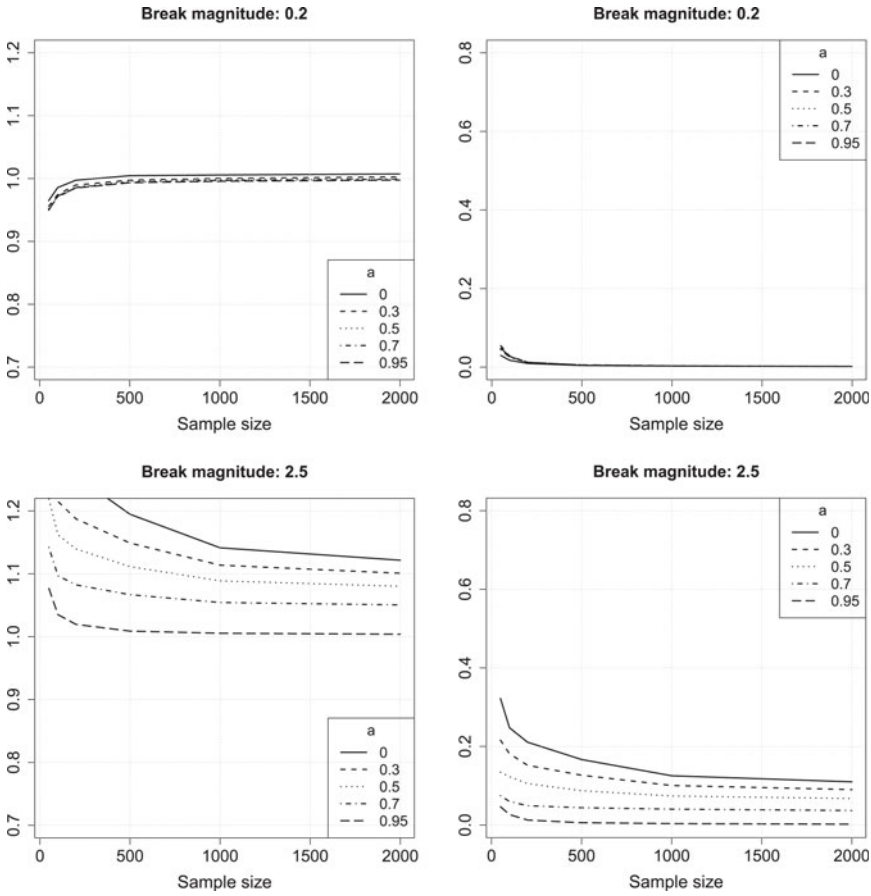

FIGURE 1. Average of residual variance (left), and variance of difference between true and LAR forecast (right) for AR(1) processes with break in mean.

correspondence between the experimental and the asymptotic values is not quite so close, and it takes some larger sample to kick in. The larger a_1, the faster the convergence.

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. Compared to scenario 1, the size of d is of minor importance. 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.

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

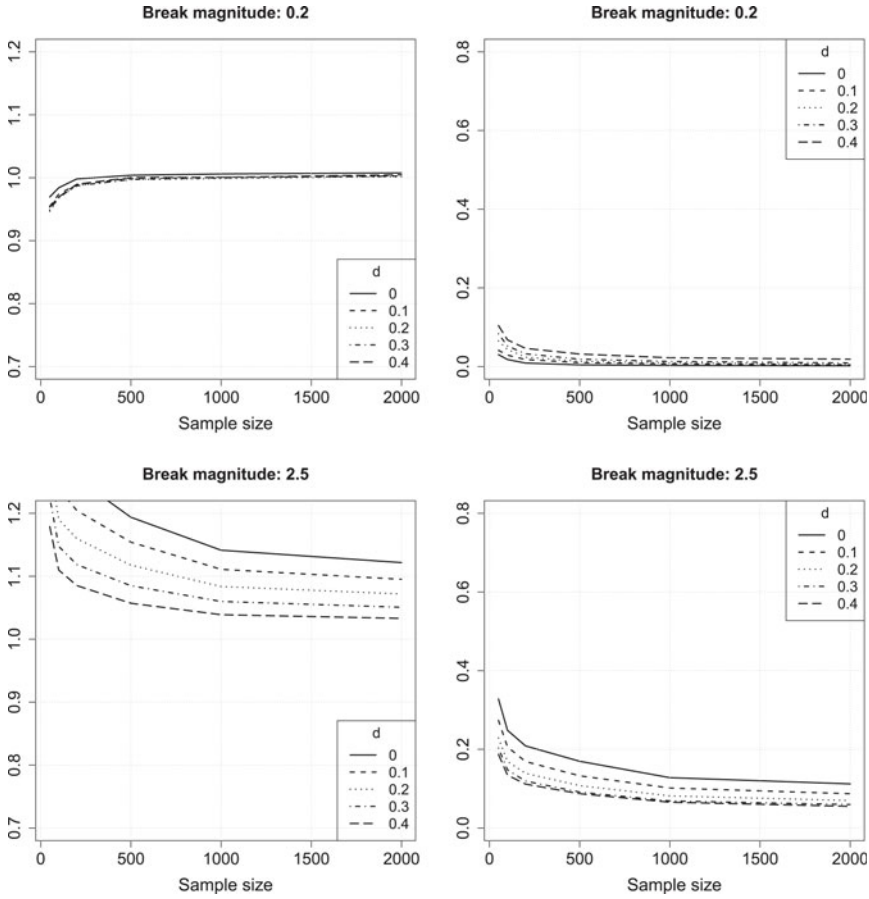

FIGURE 2. Average of residual variance (left), and variance of difference between true and LAR forecast (right) for $I(d)$ processes with break in mean.

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 4), 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: the larger the break, the stronger the bias. (E.g. for $a_1^{(1)} = 0.95$, the effect is "off the scale".) 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. To save space, we do not present the corresponding figures that are available upon request: Under long memory, the results from Scenario 3

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

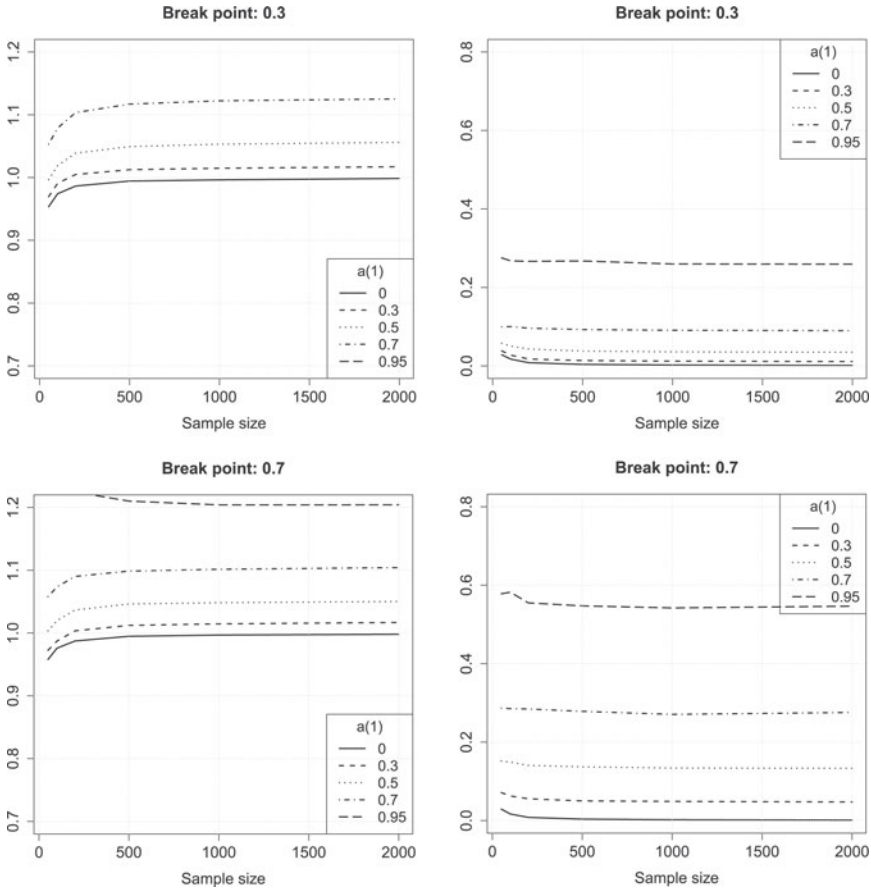

FIGURE 3. Average of residual variance (left), and variance of difference between true and LAR forecast (right) for AR(1) processes with break in dynamics.

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, Sibbertsen, and Kruse (2009) for experimental evidence.

6. CONCLUDING REMARKS

The paper considered the use of long autoregressions for forecasting processes, subject to structural change. A theoretical analysis showed that ignored 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 implicitly differencing breaks away, while the dynamics are recovered, such that the

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

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.

REFERENCES

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

Berk, K.N. (1974) Consistent autoregressive spectral estimates. *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.

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

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

Clements, M.P. & D.F. Hendry (2011) Forecasting from misspecified models in the presence of unanticipated location shifts. In M.P. Clements & 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.

Durbin, J. (1960) The fitting of time-series models. *Revue de l'Institut International de Statistique* 28(3), 233–244.

Giraitis, L., G. Kapetanios, & S. Price (2013) Adaptive forecasting in the presence of recent and ongoing structural change. *Journal of Econometrics* 177(2), 153–170.

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

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

Heinen, F., P. Sibbertsen, & R. Kruse (2009) Forecasting Long Memory Time Series Under a Break in Persistence. CREATES Research Paper 2009-53.

Lewis, R. & G.C. Reinsel (1985) Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis* 16(3), 393–411.

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

Pesaran, M.H. & A. Timmermann (2007) Selection of estimation window in the presence of breaks. *Journal of Econometrics* 137(1), 134–161.

Phillips, P.C.B. (1996) Econometric model determination. *Econometrica* 64(4), 763–812.

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.

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

Timmermann, A. & 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, & C. Hsiao (2013) Forecasting a long memory process subject to structural breaks. *Journal of Econometrics* 177(2), 171–184.

APPENDIX

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

LEMMA A.1. *Let* $\{m_t\}$ *satisfy Assumption 4 with* $1/4 < \alpha \leq 1$, *and* $\boldsymbol{m}_{t-\ell} = \left(m_{t-1}, \ldots, m_{t-\ell}\right)'$ *for some* $1 \leq \ell \leq h_T$. *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* $c_j \sim Cj^{d-1}$ *for* $0 < d < 1/2$ *as* $j \to \infty$. *Then, as* $T, h_T \to \infty$ *and* $h_T \leq CT^{\kappa}$ *for some* $\kappa < 1/3$,

1. $\sup_{1\leq\ell\leq h_T} \left\| \frac{1}{T}\sum_{t=\ell+1}^{T} \boldsymbol{m}_{t-\ell}\boldsymbol{m}'_{t-\ell} - \bar{\mu}^2 \boldsymbol{\iota}_\ell \boldsymbol{\iota}'_\ell \right\| = O\left(\frac{h_T^{1+\alpha}}{T^\alpha}\right)$ *and*
2. $\sup_{1\leq\ell\leq h_T} \left\| \frac{1}{T}\sum_{t=\ell+1}^{T} \boldsymbol{m}_{t-\ell}\boldsymbol{x}'_{t-\ell} \right\| = O_p\left(\frac{h_T}{T^{\frac{1}{2}-d}}\right)$ *when* $c_j \sim Cj^{d-1}$ *for* $0 < d < 1/2$ *or* $\sup_{1\leq\ell\leq h_T} \left\| \frac{1}{T}\sum_{t=\ell+1}^{T} \boldsymbol{m}_{t-\ell}\boldsymbol{x}'_{t-\ell} \right\| = O_p\left(\frac{h_T}{T^{1/2}}\right)$ *when* $\sum_{j\geq 0} |c_j| < \infty$.

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

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

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

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

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 (A.2) and thus negligible compared with $\left(\frac{h_T}{T}\right)^\alpha$. In the second step, we note that

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

where the maximum difference (over $1 \leq \ell \leq h_T$) between the average and the integral is of order $O\left(\max\left\{\frac{1}{T^\alpha}; \frac{\ell}{T}\right\}\right)$, thanks to the Hölder condition on ν and the fact that $0 \leq \frac{1}{T}\sum_{t=1}^{\ell} m_t^2 \leq \frac{\ell}{T}$ (again, the number of discontinuities is finite and their effect negligible, and m_t is bounded on $[0, 1]$). Summing up, Equation (A.1) holds and the desired result follows immediately.

To prove the second item, we treat ν as if it was uniformly Hölder continuous of order α, since the finite number of jumps in ν has negligible influence; see above.

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

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

elements on row k of the matrix of interest $\frac{1}{T}\sum_{t=\ell+1}^{T} \boldsymbol{m}_{t-\ell}\boldsymbol{x}'_{t-\ell}$ is

$$\sum_{j=1}^{\ell}\left|\frac{1}{T}\sum_{t=\ell+1}^{T} m_{t-k}x_{t-j}\right| \le \sum_{j=1}^{\ell}\left|\frac{1}{T}\sum_{t=\ell+1}^{T} m_{t-j}x_{t-j}\right| + \sum_{j=1}^{\ell}\left|\frac{1}{T}\sum_{t=\ell+1}^{T}\left(m_{t-k}-m_{t-j}\right)x_{t-j}\right|. \tag{A.3}$$

For the first term on the r.h.s. of (A.3), we have with the usual convention that $\sum_{m}^{n}=0$ when $m>n$ that

$$\sum_{j=1}^{\ell}\left|\frac{1}{T}\sum_{t=\ell+1}^{T} m_{t-j}x_{t-j}\right| \le \sum_{j=1}^{\ell}\left|\frac{1}{T}\sum_{t=2}^{T} m_{t-1}x_{t-1}\right| + \sum_{j=1}^{\ell}\left|\frac{1}{T}\sum_{t=2}^{\ell-j+1} m_{t-1}x_{t-1}\right| + \sum_{j=1}^{\ell}\left|\frac{1}{T}\sum_{t=T-j+2}^{T} m_{t-1}x_{t-1}\right|;$$

note that neither of the three terms on the r.h.s. of the above inequality depends on k, and thus their sum gives an upper bound for the maximum over k of the first term on the r.h.s. of (A.3). Analyzing the behavior of these three terms one by one, note that the variance of $\frac{1}{T}\sum_{t=2}^{T} m_{t-1}x_{t-1}$ – average not depending on ℓ – is easily checked to be $O\left(T^{2d-1}\right)$ thanks to the boundedness of m_t and the $O\left(j^{2d-1}\right)$ behavior of the jth-order autocovariance of x_t for $0<d<1/2$ (or absolute summability for $d=0$). At the same time,

$$\sum_{j=1}^{h_T}\left|\frac{1}{T}\sum_{t=2}^{h_T-j+1} m_{t-1}x_{t-1}\right| \le \frac{\max_{1\le t\le T}\left|m_{t-1}\right|}{T}\sum_{j=1}^{h_T}\sum_{t=2}^{h_T-j+1}\left|x_{t-1}\right| \le C\frac{\ell}{T}\sum_{t=2}^{\ell}\left|x_{t-1}\right| \le C\frac{h_T}{T}\sum_{t=2}^{h_T}\left|x_{t-1}\right|$$

for all $1\le\ell\le h_T$, such that

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

Analogously,

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

With Chebyschev's and Markov's inequalities, we thus have uniformly for $1\le\ell\le h_T$

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

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

For the second term on the r.h.s. of (A.3), we have that $\max_{1\le j,k\le \ell}\left|m_{t-k}-m_{t-j}\right| \le C\left(\frac{h_T}{T}\right)^{\alpha}$ for all $\ell \le h_T$ thanks to the Hölder condition on ν, such that

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

uniformly in ℓ. Now, the maximum over at most h_T uniformly L_2-bounded variables is of order $O_p\left(\sqrt{h_T}\right)$; by normalizing $\sum_{j=1}^{\ell}\left|\frac{1}{T}\sum_{t=\ell+1}^{T}\left(m_{t-k}-m_{t-j}\right)x_{t-j}\right|$ with $h_T\left(\frac{h_T}{T}\right)^{\alpha}T^{d-\frac{1}{2}}$ we may thus conclude that

$$\sup_{1\le \ell\le h_T}\max_{1\le k\le \ell}\sum_{j=1}^{\ell}\left|\frac{1}{T}\sum_{t=\ell+1}^{T}\left(m_{t-k}-m_{t-j}\right)x_{t-j}\right| = O_p\left(h_T^{1.5}\left(\frac{h_T}{T}\right)^{\alpha}T^{d-1/2}\right)$$

which, for $\alpha > 1/4$ and $\kappa < 1/3$, is $O_p\left(\frac{h_T}{T^{1/2-d}}\right)$ as can easily be checked (for any $\alpha > 1/4$ we have that $\frac{\alpha}{\frac{1}{2}+\alpha} > \frac{1}{3} > \kappa$ as required). Summing up, we have for the row-sum norm

$$\sup_{1\le \ell\le h_T}\max_{1\le k\le \ell}\sum_{j=1}^{\ell}\left|\frac{1}{T}\sum_{t=\ell+1}^{T}m_{t-k}x_{t-j}\right| = O_p\left(\frac{h_T}{T^{1/2-d}}\right);$$

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

$$\begin{aligned}
&\sup_{1\le \ell\le h_T}\left\|\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{m}_{t-\ell}\boldsymbol{x}_{t-\ell}'\right\|\\
&\quad\le \sqrt{\sup_{1\le \ell\le h_T}\left\|\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{m}_{t-\ell}\boldsymbol{x}_{t-\ell}'\right\|_{col}\sup_{1\le \ell\le h_T}\left\|\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{m}_{t-\ell}\boldsymbol{x}_{t-\ell}'\right\|_{row}},
\end{aligned}$$

which is $O_p\left(\frac{h_T}{T^{1/2-d}}\right)$ as required. ■

Proof of Proposition 1. We will prove that

$$\max_{1\le \ell\le h_T}\sqrt{h_T}\left\|\hat{\boldsymbol{a}}_\ell-\tilde{\boldsymbol{a}}_\ell\right\| \xrightarrow{p} 0, \tag{A.4}$$

which is slightly stronger than the required result. Obviously, (A.4) implies the convergence stated in Proposition 1; but the uniformity will further be employed in the proof of Proposition 2.

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

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

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

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

implying that

$$\left\|\Sigma_{\mu,\ell}^{-1}\right\| \le \left\|\Sigma_\ell^{-1}\right\|\left(1+\left|\frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}'_\ell\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell}\right|\left\|\boldsymbol{\iota}_\ell\boldsymbol{\iota}'_\ell\Sigma_\ell^{-1}\right\|\right).$$

Furthermore, $\boldsymbol{\iota}'_\ell\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell \ge C\ell$ has the same order as $\left\|\boldsymbol{\iota}_\ell\boldsymbol{\iota}'_\ell\Sigma_\ell^{-1}\right\| \le \left\|\boldsymbol{\iota}_\ell\boldsymbol{\iota}'_\ell\right\|\left\|\Sigma_\ell^{-1}\right\| \le C\ell \le Ch_T$, hence

$$\left\|\Sigma_{\mu,\ell}^{-1}\right\| = O\left(\left\|\Sigma_\ell^{-1}\right\|\right) = O(1) \tag{A.5}$$

uniformly in $1 \le \ell \le h_T$.

Turning our attention to the OLS estimator, we have

$$\hat{\boldsymbol{a}}_\ell = \hat{S}_\ell^{-1}\hat{G}_\ell = \hat{S}_\ell^{-1}\left(\hat{G}_\ell - \Gamma_{\mu,\ell}\right) + \left(\hat{S}_\ell^{-1} - \Sigma_{\mu,\ell}^{-1}\right)\Gamma_{\mu,\ell} + \Sigma_{\mu,\ell}^{-1}\Gamma_{\mu,\ell}$$

such that

$$\left\|\hat{\boldsymbol{a}}_\ell - \Sigma_{\mu,\ell}^{-1}\Gamma_{\mu,\ell}\right\| \le \left\|\hat{S}_\ell^{-1}\right\|\left\|\hat{G}_\ell - \Gamma_{\mu,\ell}\right\| + \left\|\hat{S}_\ell^{-1} - \Sigma_{\mu,\ell}^{-1}\right\|\left\|\Gamma_{\mu,\ell}\right\|.$$

Obviously, $\sup_{1\le\ell\le h_T}\left\|\Gamma_{\mu,\ell}\right\| = O\left(\sqrt{h_T}\right)$. We then need to analyze the remaining norms.

To do so, let us first examine

$$\hat{S}_\ell - \Sigma_{\mu,\ell} = \left(\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{x}_{t-\ell}\boldsymbol{x}'_{t-\ell} - \Sigma_\ell\right) + \left(\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{m}_{t-\ell}\boldsymbol{m}'_{t-\ell} - \bar{\mu}^2\boldsymbol{\iota}_\ell\boldsymbol{\iota}'_\ell\right) + \frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{m}_{t-\ell}\boldsymbol{x}'_{t-\ell} + \frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{x}_{t-\ell}\boldsymbol{m}'_{t-\ell}.$$

We now derive upper bounds for the norms of each of the summands on the r.h.s. of the above equation, which will give an upper bound for $\left\|\hat{S}_\ell - \Sigma_{\mu,\ell}\right\|$. It follows from Lemma A.1 that

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

Moreover, given the rate restriction $h_T = o\left(T^\kappa\right)$ for some $\kappa \le 1/4$ and the uniformly bounded variance of ε_t^2, we conclude after carefully examining the proof of Lemma 7 in Demetrescu (2009) that

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

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

and for the remaining terms, we have from Lemma A.1 that

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

Summing up,

$$\sup_{1\le\ell\le h_T}\left\|\hat{S}_\ell-\Sigma_{\mu,\ell}\right\|=O_p\left(\max\left\{\frac{h_T^2}{T};\frac{h_T}{\sqrt{T}}\right\}\right)=o_p(1).$$

Furthermore, as a consequence of this and Equation (A.5), $\left\|\hat{S}_\ell^{-1}\right\|\le\left\|\hat{S}_\ell^{-1}-\Sigma_{\mu,\ell}^{-1}\right\|+\left\|\Sigma_{\mu,\ell}^{-1}\right\|=O_p(1)$ uniformly in ℓ. Also, $\left\|\Sigma_{\mu,\ell}^{-1}\right\|\left\|\hat{S}_\ell-\Sigma_{\mu,\ell}\right\|\xrightarrow{p}0$ so $\left\|\Sigma_{\mu,\ell}^{-1}\right\|\left\|\hat{S}_\ell-\Sigma_{\mu,\ell}\right\|<1$ with probability approaching unity as $T\to\infty$ and we have in the limit that

$$\left\|\hat{S}_\ell^{-1}-\Sigma_{\mu,\ell}^{-1}\right\|\le\left\|\Sigma_{\mu,\ell}^{-1}\right\|\frac{\left\|\Sigma_{\mu,\ell}^{-1}\right\|\left\|\hat{S}_\ell-\Sigma_{\mu,\ell}\right\|}{1-\left\|\Sigma_{\mu,\ell}^{-1}\right\|\left\|\hat{S}_\ell-\Sigma_{\mu,\ell}\right\|}$$

for all ℓ. (The inequality is given explicitly by Lütkepohl, 1996, Sec. 8.4.1 11(c), but it has been used before by Lewis and Reinsel, 1985, p. 397 and Berk, 1974, eqn. 2.15.) Thus, we may conclude that

$$\sup_{1\le\ell\le h_T}\left\|\hat{S}_\ell^{-1}-\Sigma_{\mu,\ell}^{-1}\right\|=O_p\left(\sup_{1\le\ell\le h_T}\left\|\hat{S}_\ell-\Sigma_{\mu,\ell}\right\|\right)=O_p\left(\max\left\{\frac{h_T^2}{T};\frac{h_T}{\sqrt{T}}\right\}\right).$$

Similarly,

$$\hat{G}_\ell-\Gamma_{\mu,\ell}=\left(\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{x}_{t-\ell}x_t-\Gamma_\ell\right)+\left(\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{m}_{t-\ell}m_t-\bar{\mu}^2\boldsymbol{\iota}_\ell\right)+\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{x}_{t-\ell}m_t+\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{m}_{t-\ell}x_t.$$

We have analogously to the relations above that

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

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

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

$$\sup_{1\le\ell\le h_T}\left\|\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{x}_{t-\ell}x_t-\Gamma_\ell\right\|=O_p\left(\sqrt{\frac{h_T}{T}}\right)$$

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

such that, summing up,

$$\sup_{1\le\ell\le h_T}\left\|\hat{G}_\ell-\Gamma_{\mu,\ell}\right\|=O_p\left(\max\left\{\frac{h_T^{1.5}}{T};\sqrt{\frac{h_T}{T}}\right\}\right).$$

Hence

$$\sup_{1\le\ell\le h_T}\left\|\hat{\boldsymbol{a}}_\ell-\Sigma_{\mu,\ell}^{-1}\Gamma_{\mu,\ell}\right\|=O_p\left(\sqrt{h_T}\cdot\max\left\{\frac{h_T^2}{T};\frac{h_T}{\sqrt{T}}\right\}\right)+O_p\left(\max\left\{\frac{h_T^{1.5}}{T};\sqrt{\frac{h_T}{T}}\right\}\right).$$

This is $o_p\left(h_T^{-1/2}\right)$ when suitably choosing $\kappa\le 1/4$, as required for (A.4).

Focusing now on the "centering" sequence

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

use the Sherman–Morrison formula again to obtain that

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

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

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

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

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

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

Since $\sum_{j=1}^{h_T} \tilde{a}_{j,h_T} \to 1$ as $T \to \infty$ by Proposition 1 (see also Remark 2), 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

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

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 $\left(1 - \boldsymbol{\iota}'_{h_T} \boldsymbol{a}_{h_T}\right)$ is bounded, see above and

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

since $\operatorname{Var}\left(\boldsymbol{x}'_{t-h_T} \Sigma_{h_T}^{-1} \boldsymbol{\iota}_{h_T}\right) = \boldsymbol{\iota}'_{h_T} \Sigma_{h_T}^{-1} \operatorname{Cov}\left(\boldsymbol{x}_{t-h_T}\right) \Sigma_{h_T}^{-1} \boldsymbol{\iota}_{h_T} = \boldsymbol{\iota}'_{h_T} \Sigma_{h_T}^{-1} \boldsymbol{\iota}_{h_T}$. At the same time,

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

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

Proof of Proposition 2. For a fitted model or order ℓ, the residuals are

$$\begin{aligned}\hat{\varepsilon}_{t,\ell} &= y_t - \hat{\boldsymbol{a}}'_\ell \boldsymbol{y}_{t-\ell} = x_t - \hat{\boldsymbol{a}}'_\ell \boldsymbol{x}_{t-\ell} + m_t - \hat{\boldsymbol{a}}'_\ell \boldsymbol{m}_{t-\ell} \\ &= \varepsilon_t - \left(\hat{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)' \boldsymbol{x}_{t-\ell} + m_t - \hat{\boldsymbol{a}}'_\ell \boldsymbol{m}_{t-\ell}.\end{aligned}$$

We start with the case $\ell \geq p$, where

$$\begin{aligned}\hat{\sigma}_\ell^2 = \frac{1}{T} \sum_{t=\ell+1}^{T} \hat{\varepsilon}_{t,\ell}^2 &= \frac{1}{T} \sum_{t=\ell+1}^{T} \varepsilon_t^2 + \frac{2}{T} \sum_{t=\ell+1}^{T} \varepsilon_t \left(-\left(\hat{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)' \boldsymbol{x}_{t-\ell} + m_t - \hat{\boldsymbol{a}}'_\ell \boldsymbol{m}_{t-\ell}\right) \\ &\quad + \frac{1}{T} \sum_{t=\ell+1}^{T} \left(-\left(\hat{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)' \boldsymbol{x}_{t-\ell} + m_t - \hat{\boldsymbol{a}}'_\ell \boldsymbol{m}_{t-\ell}\right)^2 \\ &= A_T + B_T + C_T.\end{aligned}$$

Analyzing the three terms in turn, we have first

$$A_T = \frac{1}{T} \sum_{t=\ell+1}^{T} \varepsilon_t^2 = \frac{1}{T} \sum_{t=h_T+1}^{T} \varepsilon_t^2 - \frac{1}{T} \sum_{t=l+1}^{h_T} \varepsilon_t^2$$

where, for all $p \leq \ell \leq h_T$,

$$0 \leq \frac{1}{T} \sum_{t=l+1}^{h_T} \varepsilon_t^2 \leq \frac{1}{T} \sum_{t=1}^{h_T} \varepsilon_t^2 = O_p\left(\frac{h_T}{T}\right)$$

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

such that

$$\max_{p\le\ell\le h_T}\left|\frac{1}{T}\sum_{t=\ell+1}^{T}\varepsilon_t^2-\frac{1}{T}\sum_{t=h_T+1}^{T}\varepsilon_t^2\right|=O_p\left(\frac{h_T}{T}\right);$$

moreover,

$$\frac{1}{T}\sum_{t=h_T+1}^{T}\varepsilon_t^2=\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2+O_p\left(\frac{h_T}{T}\right)=\sigma^2+o_p(1)$$

with the $o_p(1)$ term obviously not depending on ℓ, so $\max_{p\le\ell\le h_T}\left|A_T-\sigma^2\right|=o_p(1)$.

Second, for examining

$$B_T=-\frac{2}{T}\left(\hat{\boldsymbol{a}}_\ell-\boldsymbol{a}_\ell\right)'\sum_{t=\ell+1}^{T}\varepsilon_t\boldsymbol{x}_{t-\ell}+\frac{2}{T}\sum_{t=\ell+1}^{T}\varepsilon_t\left(m_t-\hat{\boldsymbol{a}}_\ell'\boldsymbol{m}_{t-\ell}\right)=B_{T1}+B_{T2},$$

note that

$$\begin{aligned}\left\|\hat{\boldsymbol{a}}_\ell-\boldsymbol{a}_\ell\right\|&\le\left\|\hat{\boldsymbol{a}}_\ell-\tilde{\boldsymbol{a}}_\ell\right\|+\left\|\tilde{\boldsymbol{a}}_\ell-\boldsymbol{a}_\ell\right\|\\&=o_p(1)+\left\|\frac{\bar{\mu}^2}{1+\bar{\mu}^2\boldsymbol{\iota}_\ell'\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell}\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell\left(1-\boldsymbol{\iota}_\ell'\boldsymbol{a}_\ell\right)\right\|.\end{aligned}$$

The $o_p(1)$ term on the r.h.s. is uniform in $p\le\ell\le h_T$, see Equation (A.4), while the norm converges as $\ell\to\infty$ (see e.g. the discussion preceding Equation (A.5)) and is thus bounded. Hence $\left\|\hat{\boldsymbol{a}}_\ell-\boldsymbol{a}_\ell\right\|=O_p(1)$ uniformly in $p\le\ell\le h_T$. Note now that

$$\begin{aligned}\left\|\sum_{t=h_T+1}^{T}\varepsilon_t\boldsymbol{x}_{t-\ell}\right\|&=\sqrt{\sum_{i=1}^{\ell}\left(\sum_{t=h_T+1}^{T}\varepsilon_t x_{t-i}\right)^2}\le\sqrt{\sum_{i=1}^{h_T}\left(\sum_{t=h_T+1}^{T}\varepsilon_t x_{t-i}\right)^2}\\&=\left\|\sum_{t=h_T+1}^{T}\varepsilon_t\boldsymbol{x}_{t-h_T}\right\|\end{aligned}$$

and similarly $\left\|\sum_{t=\ell+1}^{h_T}\varepsilon_t\boldsymbol{x}_{t-\ell}\right\|\le\left\|\sum_{t=1}^{h_T}\varepsilon_t\boldsymbol{x}_{t-h_T}\right\|$ such that it follows for any $p\le\ell\le h_T$ that

$$\begin{aligned}\left\|\frac{1}{T}\sum_{t=\ell+1}^{T}\varepsilon_t\boldsymbol{x}_{t-\ell}\right\|&\le\left\|\frac{1}{T}\sum_{t=h_T+1}^{T}\varepsilon_t\boldsymbol{x}_{t-\ell}\right\|+\left\|\frac{1}{T}\sum_{t=\ell+1}^{h_T}\varepsilon_t\boldsymbol{x}_{t-\ell}\right\|\\&\le\left\|\frac{1}{T}\sum_{t=h_T+1}^{T}\varepsilon_t\boldsymbol{x}_{t-h_T}\right\|+\left\|\frac{1}{T}\sum_{t=\ell+1}^{h_T}\varepsilon_t\boldsymbol{x}_{t-h_T}\right\|,\end{aligned}$$

where in turn $\left\|\frac{1}{T}\sum_{t=h_T+1}^{T}\varepsilon_t\boldsymbol{x}_{t-h_T}\right\|=O_p\left(\sqrt{\frac{h_T}{T}}\right)$ and $\left\|\frac{1}{T}\sum_{t=\ell+1}^{h_T}\varepsilon_t\boldsymbol{x}_{t-h_T}\right\|=O_p\left(\frac{h_T^2}{T}\right)$ uniformly in ℓ, leading to

$$\max_{p\le\ell\le h_T}|B_{T1}|\le\max_{p\le\ell\le h_T}\left\|\hat{\boldsymbol{a}}_\ell-\boldsymbol{a}_\ell\right\|\max_{p\le\ell\le h_T}\left\|\frac{1}{T}\sum_{t=\ell+1}^{T}\varepsilon_t\boldsymbol{x}_{t-\ell}\right\|=o_p(1).$$

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

For examining B_{T2}, note that $\|\hat{\boldsymbol{a}}_\ell\| = \|\boldsymbol{a}_\ell\| + o_p\left(\ell^{-1/2}\right) = O_p(1)$ and proceeds as above to obtain uniform (in ℓ) bounds for $\frac{2}{T}\sum_{t=\ell+1}^{T}\varepsilon_t m_t$ and $\frac{2}{T}\sum_{t=\ell+1}^{T}\varepsilon_t \boldsymbol{m}_{t-\ell}$ leading after some algebra to

$$B_{T2} = \frac{2}{T}\sum_{t=\ell+1}^{T}\varepsilon_t m_t - \hat{\boldsymbol{a}}_\ell'\frac{2}{T}\sum_{t=\ell+1}^{T}\varepsilon_t \boldsymbol{m}_{t-\ell} = O_p\left(T^{-1/2}\right) + O_p\left(\sqrt{\frac{h_T}{T}}\right)$$

uniformly in $p \le \ell \le h_T$, or $\max_{p\le\ell\le h_T}|B_{T2}| = o_p(1)$.

Third,

$$\begin{aligned} C_T &= \frac{1}{T}\sum_{t=\ell+1}^{T}\left(\left(\hat{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)'\boldsymbol{x}_{t-\ell}\right)^2 + \frac{1}{T}\sum_{t=\ell+1}^{T}\left(m_t - \hat{\boldsymbol{a}}_\ell'\boldsymbol{m}_{t-\ell}\right)^2 \\ &\quad -\frac{2}{T}\sum_{t=\ell+1}^{T}\left(\left(\hat{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)'\boldsymbol{x}_{t-\ell}\right)\left(m_t - \hat{\boldsymbol{a}}_\ell'\boldsymbol{m}_{t-\ell}\right). \\ &= C_{T1} + C_{T2} + C_{T3}. \end{aligned}$$

Examining C_{T1}, write $\hat{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell = \hat{\boldsymbol{a}}_\ell - \tilde{\boldsymbol{a}}_\ell + \tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell$ to obtain

$$\begin{aligned} C_{T1} &= \frac{1}{T}\sum_{t=\ell+1}^{T}\left(\left(\hat{\boldsymbol{a}}_\ell - \tilde{\boldsymbol{a}}_\ell\right)'\boldsymbol{x}_{t-\ell}\right)^2 + \frac{1}{T}\sum_{t=\ell+1}^{T}\left(\left(\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)'\boldsymbol{x}_{t-\ell}\right)^2 \\ &\quad + \frac{2}{T}\sum_{t=\ell+1}^{T}\left(\left(\hat{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)'\boldsymbol{x}_{t-\ell}\right)\left(\left(\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)'\boldsymbol{x}_{t-\ell}\right). \end{aligned}$$

For the first term on the r.h.s. of the above equation, we have

$$\frac{1}{T}\sum_{t=\ell+1}^{T}\left(\left(\hat{\boldsymbol{a}}_\ell - \tilde{\boldsymbol{a}}_\ell\right)'\boldsymbol{x}_{t-\ell}\right)^2 = \left(\hat{\boldsymbol{a}}_\ell - \tilde{\boldsymbol{a}}_\ell\right)'\left(\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{x}_{t-\ell}\boldsymbol{x}_{t-\ell}'\right)\left(\hat{\boldsymbol{a}}_\ell - \tilde{\boldsymbol{a}}_\ell\right)$$

where $\|\hat{\boldsymbol{a}}_\ell - \tilde{\boldsymbol{a}}_\ell\| \to 0$ uniformly in $p \le \ell \le h_T$, see Equation (A.4), and the norm of the sample covariance matrix is easily shown to be uniformly (in ℓ) bounded in probability such that

$$\max_{p\le\ell\le h_T}\frac{1}{T}\sum_{t=\ell+1}^{T}\left(\left(\hat{\boldsymbol{a}}_\ell - \tilde{\boldsymbol{a}}_\ell\right)'\boldsymbol{x}_{t-\ell}\right)^2 = o_p(1).$$

The second term on the r.h.s. gives

$$\begin{aligned} \frac{1}{T}\sum_{t=\ell+1}^{T}\left(\left(\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)'\boldsymbol{x}_{t-\ell}\right)^2 &= \left(\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)'\Sigma_\ell\left(\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right) \\ &\quad + \left(\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right)'\left(\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{x}_{t-\ell}\boldsymbol{x}_{t-\ell}' - \Sigma_\ell\right)\left(\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\right). \end{aligned}$$

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

With $\|\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell\| = O_p(1)\ \forall\ p \le \ell \le h_T$ as above and $\left\| \frac{1}{T}\sum_{t=\ell+1}^{T} \boldsymbol{x}_{t-\ell}\boldsymbol{x}'_{t-\ell} - \Sigma_\ell \right\| \xrightarrow{p} 0$ $\forall\ p \le \ell \le h_T$, it follows for the second term on the r.h.s. that

$$\frac{1}{T}\sum_{t=\ell+1}^{T}\left((\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell)'\boldsymbol{x}_{t-\ell}\right)^2 = (\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell)'\,\Sigma_\ell\,(\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell) + o_p(1)$$

uniformly in $p \le \ell \le h_T$, or

$$\max_{p\le\ell\le h_T}\left|\frac{1}{T}\sum_{t=\ell+1}^{T}\left((\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell)'\boldsymbol{x}_{t-\ell}\right)^2 - \frac{\bar{\mu}^4\left(1-\boldsymbol{\iota}'_\ell\boldsymbol{a}_\ell\right)^2\boldsymbol{\iota}'_\ell\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell}{\left(1+\bar{\mu}^2\boldsymbol{\iota}'_\ell\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell\right)^2}\right| = o_p(1).$$

The absolute value of the third term on the r.h.s. can be bounded via the Cauchy–Schwarz inequality,

$$\left|\sum_{t=\ell+1}^{T}\left((\hat{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell)'\boldsymbol{x}_{t-\ell}\right)\left((\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell)'\boldsymbol{x}_{t-\ell}\right)\right| \le \sqrt{\sum_{t=\ell+1}^{T}\left((\hat{\boldsymbol{a}}_\ell - \tilde{\boldsymbol{a}}_\ell)'\boldsymbol{x}_{t-\ell}\right)^2\sum_{t=\ell+1}^{T}\left((\tilde{\boldsymbol{a}}_\ell - \boldsymbol{a}_\ell)'\boldsymbol{x}_{t-\ell}\right)^2}$$

so it also vanishes uniformly. Thus,

$$\max_{p\le\ell\le h_T}\left|C_{T1} - \frac{\bar{\mu}^4\left(1-\boldsymbol{\iota}'_\ell\boldsymbol{a}_\ell\right)^2\boldsymbol{\iota}'_\ell\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell}{\left(1+\bar{\mu}^2\boldsymbol{\iota}'_\ell\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell\right)^2}\right| = o_p(1).$$

For C_{T2}, we have that $m_t - \hat{\boldsymbol{a}}'_\ell\boldsymbol{m}_{t-\ell} = m_t\left(1-\hat{\boldsymbol{a}}'_\ell\boldsymbol{\iota}_\ell\right)$ for $\ell \le t \le \tau T$ and $\tau T + h_T \le t \le T$, and otherwise $\max_{p\le\ell\le h_T}\left|m_t - \hat{\boldsymbol{a}}'_\ell\boldsymbol{m}_{t-\ell}\right| = O_p(h_T)$, since $\left\|\hat{\boldsymbol{a}}_\ell\right\| = o_p(1)$ as above, so

$$\frac{1}{T}\sum_{t=\ell+1}^{T}\left(m_t - \hat{\boldsymbol{a}}'_\ell\boldsymbol{m}_{t-\ell}\right)^2 = \left(1-\hat{\boldsymbol{a}}'_\ell\boldsymbol{\iota}_\ell\right)\left(\frac{1}{T}\sum_{t=\ell+1}^{T}m_t^2 - \frac{1}{T}\sum_{t=\tau T-\ell+1}^{\tau T}m_t^2\right) + O_p\left(\frac{h_T^2}{T}\right)$$

uniformly. Then, with $\frac{1}{T}\sum_{t=1}^{T}m_t^2 \to \bar{\mu}^2$, $\max_{p\le\ell\le h_T}\left|\frac{1}{T}\sum_{t=\tau T-\ell+1}^{\tau T}m_t^2\right| = O\left(\frac{h_T}{T}\right)$ and

$$\hat{\boldsymbol{a}}'_\ell\boldsymbol{\iota}_\ell = \tilde{\boldsymbol{a}}'_\ell\boldsymbol{\iota}_\ell + \left(\hat{\boldsymbol{a}}_\ell - \tilde{\boldsymbol{a}}_\ell\right)'\boldsymbol{\iota}_\ell = 1 + o_p(1)$$

uniformly in $p \le \ell \le h_T$, it follows that

$$\max_{p\le\ell j\le h_T}\left|\frac{1}{T}\sum_{t=\ell+1}^{T}\left(m_t - \hat{\boldsymbol{a}}'_\ell\boldsymbol{m}_{t-\ell}\right)^2\right| = o_p(1).$$

Using Cauchy–Schwarz again, we have $|C_{T3}| \le \sqrt{C_{T1}C_{T2}} = o_p(1)$ uniformly in $p \le \ell \le h_T$.

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

Summing up, we have that

$$\max_{p\le\ell\le h_T}\left|\hat{\sigma}_\ell^2-\sigma^2-\frac{\bar{\mu}^4\left(1-\boldsymbol{\iota}_\ell'\boldsymbol{a}_\ell\right)^2\boldsymbol{\iota}_\ell'\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell}{\left(1+\bar{\mu}^2\boldsymbol{\iota}_\ell'\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell\right)^2}\right|\xrightarrow{p}0.$$

Since $AIC(\ell)=\ln\hat{\sigma}_\ell^2+\frac{2\ell}{T}$ and $\ell\le h_T$, this implies in turn

$$\max_{p\le\ell\le h_T}\left|AIC(\ell)-\ln\left(\sigma^2+\frac{\bar{\mu}^4\left(1-\boldsymbol{\iota}_\ell'\boldsymbol{a}_\ell\right)^2\boldsymbol{\iota}_\ell'\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell}{\left(1+\bar{\mu}^2\boldsymbol{\iota}_\ell'\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell\right)^2}\right)\right|\xrightarrow{p}0.$$

The logarithm in the equation above is a sequence converging to $\ln\sigma^2$ from above. Hence, for any fixed ℓ, the probability to observe an $AIC(\ell)$ smaller than *all* $AIC(k)$ for $\ell<k\le h_T$ converges to zero, implying that $\underset{p\le\ell\le h_T}{\arg\min}\,AIC(\ell)\xrightarrow{p}\infty$ as required.

Since the true model order is equal to p, it is easily shown that a lag length smaller than p will not be chosen asymptotically; we omit the details. The proof is complete. ■

Proof of Proposition 3. The steps of the proof are essentially the same as in the proof of Proposition 1, and we use the same notation with $\tilde{\boldsymbol{a}}_\ell=\Sigma_{\mu,\ell}^{-1}\Gamma_{\mu,\ell}$ for each $1\le\ell\le h_T$ etc.

Hassler and Kokoszka (2010) show that, if $j^{1-d}b_j\to0$ 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 $0<d<\frac{1}{2}$, and on the analogous results for absolutely summable coefficients for $d=0$.

For $0<d<1/2$, $\gamma_j=\mathrm{Cov}\left(x_t,x_{t-j}\right)=O\left(j^{2d-1}\right)$ such that $\|\Sigma_\ell\|\le C\ell^{2d}\le Ch_T^{2d}$ and $\|\Gamma_\ell\|\le Ch_T^d$. Still, $\left\|\Sigma_\ell^{-1}\right\|=O(1)$ uniformly in ℓ like in the short-memory case (which is recovered for $d=0$). Also, $\boldsymbol{\iota}_\ell\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell'\ge C\ell^{1-2d}$ since Σ_ℓ^{-1} is positive definite and its smallest eigenvalue is at least of order ℓ^{-2d}. It then follows, like in the proof of Proposition 1, that

$$\left\|\Sigma_{\mu,\ell}^{-1}\right\|=O\left(\left\|\Sigma_\ell^{-1}\right\|\frac{\left\|\boldsymbol{\iota}_\ell\boldsymbol{\iota}_\ell'\Sigma_\ell^{-1}\right\|}{\boldsymbol{\iota}_\ell'\Sigma_\ell^{-1}\boldsymbol{\iota}_\ell}\right)=O\left(\ell^{2d}\right)=O\left(h_T^{2d}\right)$$

for all $1\le\ell\le h_T$. Therefore,

$$\begin{aligned}\sup_{1\le\ell\le h_T}\left\|\hat{\boldsymbol{a}}_\ell-\Sigma_{\mu,\ell}^{-1}\Gamma_{\mu,\ell}\right\|\le{}&\sup_{1\le\ell\le h_T}\left\|\hat{S}_\ell^{-1}\right\|\sup_{1\le\ell\le h_T}\left\|\hat{G}_\ell-\Gamma_{\mu,\ell}\right\|\\&+\sup_{1\le\ell\le h_T}\left\|\hat{S}_\ell^{-1}-\Sigma_{\mu,\ell}^{-1}\right\|\sup_{1\le\ell\le h_T}\left\|\Gamma_{\mu,\ell}\right\|\end{aligned}\tag{A.6}$$

where $\left\|\Gamma_{\mu,\ell}\right\|\le\|\Gamma_\ell\|+\bar{\mu}^2\|\boldsymbol{\iota}_\ell\|\le C\sqrt{h_T}$ for all $1\le\ell\le h_T$.

To establish the behavior of the r.h.s. of the inequality (A.6), the same norms as in the proof of Proposition 1 need to be examined.

From Poskitt (2007, Thm. 1), it follows that

$$\sup_{1\le\ell\le h_T}\left\|\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{x}_{t-\ell}\boldsymbol{x}_{t-\ell}'-\Sigma_\ell\right\|=O_p\left(h_T\left(\frac{\log T}{T}\right)^{1/2-d}\right),$$

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

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

$$\sup_{1\le\ell\le h_T}\left\|\hat{S}_\ell-\Sigma_{\mu,\ell}\right\|=O_p\left(\max\left\{\frac{h_T^{1+\alpha}}{T^\alpha};h_T\left(\frac{\log T}{T}\right)^{1/2-d}\right\}\right).$$

With both $h_T\left(\frac{\log T}{T}\right)^{1/2-d}$ vanishing and $\frac{h_T^{1+\alpha}}{T^\alpha}$ dominated by h_T^{2d} since $\kappa<\frac{\alpha}{2+\alpha+4d}$ and $\frac{\alpha}{2+\alpha+4d}<\frac{\alpha}{1+\alpha-2d}$, we further have that

$$\sup_{1\le\ell\le h_T}\left\|\hat{S}_\ell^{-1}-\Sigma_{\mu,\ell}^{-1}\right\|=O_p\left(h_T^{4d}\cdot\max\left\{\frac{h_T^{1+\alpha}}{T^\alpha};h_T\left(\frac{\log T}{T}\right)^{1/2-d}\right\}\right)=o_p(1)$$

and hence

$$\begin{aligned}\sup_{1\le\ell\le h_T}\left\|\hat{S}_\ell^{-1}\right\|&\le\sup_{1\le\ell\le h_T}\left\|\hat{S}_\ell^{-1}-\Sigma_{\mu,\ell}^{-1}\right\|+\sup_{1\le\ell\le h_T}\left\|\Sigma_{\mu,\ell}^{-1}\right\|\\&=O_p\left(\sup_{1\le\ell\le h_T}\left\|\Sigma_{\mu,\ell}^{-1}\right\|\right)=O_p\left(h_T^{2d}\right).\end{aligned}$$

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

$$\sup_{1\le\ell\le h_T}\left\|\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{x}_{t-\ell}m_t\right\|=O_p\left(\frac{\sqrt{h_T}}{T^{\frac{1}{2}-d}}\right)=\sup_{1\le\ell\le h_T}\left\|\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{m}_{t-\ell}x_t\right\|.$$

Lemma A.1 further allows us to conclude that

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

and, using again Poskitt (2007, Thm. 1), we have that

$$\sup_{1\le\ell\le h_T}\left\|\frac{1}{T}\sum_{t=\ell+1}^{T}\boldsymbol{x}_{t-\ell}x_t-\Gamma_\ell\right\|=O_p\left(\sqrt{h_T}\left(\frac{\log T}{T}\right)^{1/2-d}\right).$$

Hence

$$\sup_{1\le\ell\le h_T}\left\|\hat{G}_\ell-\Gamma_{\mu,\ell}\right\|=O_p\left(\max\left\{\frac{h_T^{1/2+\alpha}}{T^\alpha};\sqrt{h_T}\left(\frac{\log T}{T}\right)^{1/2-d}\right\}\right)$$

such that $\sup_{1\le\ell\le h_T}\left\|\hat{\boldsymbol{a}}_\ell-\Sigma_{\mu,\ell}^{-1}\Gamma_{\mu,\ell}\right\|$ is

$$\begin{aligned}&O_p\left(h_T^{2d}\cdot\max\left\{\frac{h_T^{1/2+\alpha}}{T^\alpha};\sqrt{h_T}\left(\frac{\log T}{T}\right)^{1/2-d}\right\}\right)\\&\quad+O_p\left(\sqrt{h_T}\cdot\max\left\{\frac{h_T^{1+\alpha+4d}}{T^\alpha};h_T^{1+4d}\left(\frac{\log T}{T}\right)^{1/2-d}\right\}\right),\end{aligned}$$

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

i.e., $\sup_{1\le\ell\le h_T}\left\|\hat{\boldsymbol{a}}_\ell-\Sigma_{\mu,\ell}^{-1}\Gamma_{\mu,\ell}\right\| = O_p\left(h_T^{-1/2}\max\left\{\frac{h_T^{2+\alpha+4d}}{T^\alpha};h_T^{2+4d}\left(\frac{\log T}{T}\right)^{1/2-d}\right\}\right)$. To obtain the desired convergence rate for $\left\|\hat{\boldsymbol{a}}_\ell-\Sigma_{\mu,\ell}^{-1}\Gamma_{\mu,\ell}\right\|$, it suffices to show that

$$\max\left\{\frac{T^{\kappa(2+\alpha+4d)}}{T^\alpha};(\log T)^{1/2-d}\,\frac{T^{\kappa(2+4d)}}{T^{1/2-d}}\right\}\to 0,$$

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

Proof of Corollary 2. To show that Corollary 1 still holds under the assumptions of Proposition 3, note that

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

since the coefficients $\tilde{a}_{j,h_T}$ sum up to 1 whenever $\boldsymbol{\iota}_{h_T}\Sigma_{h_T}^{-1}\boldsymbol{\iota}_{h_T}'\to\infty$ (see the proof of Corollary 1) and indeed $\boldsymbol{\iota}_{h_T}\Sigma_{h_T}^{-1}\boldsymbol{\iota}_{h_T}'\ge Ch_T^{1-2d}\to\infty$ as argued above. At the same time,

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

since $\left\|\hat{\boldsymbol{a}}_{h_T}-\tilde{\boldsymbol{a}}_{h_T}\right\|=o\left(h_T^{-1/2}\right)$ and $\left\|\boldsymbol{m}_{T+1-h_T}\right\|=O\left(\sqrt{h_T}\right)=\left\|\boldsymbol{x}_{T+1-h_T}\right\|$ thanks to the boundedness of m_t and the uniformly bounded variance of x_t. Then,

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

The first term is easily shown to vanish, since

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

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 A.1). For the second term, note that $x_{T+1}-\sum_{j=1}^{\infty}a_jx_{T+1-j}$ is

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press

the forecast error from a projection of x_t on its infinite past, and $x_{T+1} - \boldsymbol{x}'_{T+1-h_T} \Sigma_{h_T}^{-1} \Gamma_{h_T}$ is 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 $\boldsymbol{\iota}_{h_T} \Sigma_{h_T}^{-1} \boldsymbol{\iota}'_{h_T} \to \infty$. Hence, the proof is complete. ■

https://doi.org/10.1017/S0266466615000225 Published online by Cambridge University Press