Learning generates Long Memory

Guillaume Chevillon, Sophocles Mavroeidis

▶ **To cite this version:**

Guillaume Chevillon, Sophocles Mavroeidis. Learning generates Long Memory. 2011, pp.57. hal-00661012v1

HAL Id: hal-00661012

https://essec.hal.science/hal-00661012v1

Submitted on 18 Jan 2012 (v1), last revised 15 Oct 2013 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Learning generates Long Memory

Research Center
ESSEC Working Paper 1113
Novembre 2011

Guillaume Chevillon

Sophocles Mavroeidis

ESSEC
BUSINESS SCHOOL

Learning generates Long Memory*

Guillaume Chevillon
ESSEC Business School
& CREST-INSEE, Paris

Sophocles Mavroeidis
University of Oxford

November 24, 2011

Abstract

We consider a prototypical representative-agent forward-looking model, and study the low frequency variability of the data when the agent's beliefs about the model are updated through linear learning algorithms. We find that learning in this context can generate strong persistence. The degree of persistence depends on the weights agents place on past observations when they update their beliefs, and on the magnitude of the feedback from expectations to the endogenous variable. When the learning algorithm is recursive least squares, long memory arises when the coefficient on expectations is sufficiently large. In algorithms with discounting, long memory provides a very good approximation to the low-frequency variability of the data. Hence long memory arises endogenously, due to the self-referential nature of the model, without any persistence in the exogenous shocks. This is distinctly different from the case of rational expectations, where the memory of the endogenous variable is determined exogenously. Finally, this property of learning is used to shed light on some well-known empirical puzzles.

JEL Codes: C1, E3;

Keywords: Learning, Long Memory, Persistence, Present-Value Models.

*Emails for correspondance: chevillon@essec.edu and sophocles.mavroeidis@economics.ox.ac.uk . The authors would like to thank Richard Baillie, Peter Howitt, Rustam Ibragimov, Frank Kleibergen, Guy Laroque, Ulrich Müller, Mark Watson, Ken West, as well as the participants in the 2010 NBER summer institute for helpful comments and discussions. We also benefited from comments received at the Nordic Econometric Meeting in Lund, the Netherlands Econometric Study Group in Leuven, the ESEM in Oslo, EC^2 in Toulouse as well as from seminar participants at Cambridge, CREST, ESSEC, GREQAM, Nottingham, Oxford and Rotterdam.

1 Introduction

In many economic models, the behavior of economic agents depends on their expectations of the current or future states of the economy. For example, in the new Keynesian policy model, prices are set according to firms' expectations of future marginal costs, consumption is determined according to consumers' expectations of future income, and policy makers' actions depend on their expectations of the current and future macroeconomic conditions, see Clarida, Galí and Gertler (1999). In asset pricing models, prices are determined by expected dividends and future price appreciation, see Campbell and Shiller (1987).

In a rational expectations equilibrium, these models imply that the dynamics of the endogenous variables are determined exogenously and therefore, these models typically fail to explain the observed persistence in the data. It has long been recognized that bounded rationality, or learning, may induce richer dynamics and can account for some of the persistence in the data, see Sargent (1993) and Evans and Honkapohja (2009). In a related paper, Chevillon, Massmann and Mavroeidis (2010) showed that the persistence induced by learning can be so strong as to invalidate conventional econometric methods of estimation and inference.

In this paper, we explore the issue of persistence further, with particular emphasis on the impact of the memory of the learning algorithm on the dynamics of the endogenous variable at low frequencies. Specifically, we characterize learning algorithms in terms of the effective length of the sample of past data that agents use to update their beliefs. 'Short window' learning corresponds to the case when past observations are heavily discounted, a classic example being exponentially weighted moving averaging, also known as 'constant gain least squares' or CGLS. This is commonly referred to as 'perpetual learning' and is very popular in empirical work, see Evans and Honkapohja (2009). 'Long window' learning corresponds to mild or no discounting of past observations, such as recursive least squares.

When we add learning to a prototypical forward-looking model, we find that the resulting dynamics of the endogenous variable exhibit long range dependence. We measure the degree of long range dependence, or the memory of the process, in terms of the order of magnitude of the variance of partial sums of the process. In stationary cases, we also study the behavior of the spectrum near zero and the autocorrelation function at long lags. We find that the memory of the process depends on both the length of the learning algorithm, and the feedback that

expectations have on the process. The latter is governed by the coefficient on expectations, which in many applications is interpretable as a discount factor. It is important to stress that this coefficient plays no role for the memory of the process under rational expectations. Finally, these results are established under the assumption that exogenous shocks have short memory, and hence, it is shown that long memory can arise completely endogenously through learning.

The above results provide a structural interpretation of a phenomenon which has been found to be important for many economic time series. The other main explanations of long-range dependence that we are aware of are: (i) aggregation of short memory series — either cross-sectionally (with beta-distributed weights in Granger, 1980, or with heterogeneity in Abadir and Talmain, 2002 and Zaffaroni, 2004) or temporally across mixed-frequencies (Chambers, 1998); (ii) occasional breaks that can produce fractional integration (Parke, 1999) or be mistaken for it (Granger and Ding, 1996, Diebold and Inoue, 2001, or Perron and Qu, 2007); and (iii) some form of nonlinearity (see e.g. Davidson and Sibbertsen, 2005, and Miller and Park, 2010). Ours is (to our knowledge) the first explanation that traces the source of long-range dependence to the behavior of agents, and the self-referential nature of economic outcomes.

The paper is organized as follows. Section 2 presents the modelling framework and introduces the concept of length of the learning window that is central to our analysis. We then present in section 3 our analytical results regarding the relation between the length of the learning algorithm and the long memory properties of the data. Monte Carlo simulation evidence confirming our theoretical predictions follows in section 4. Finally, section 5 considers the implications of adaptive learning in the present values models of Campbell and Shiller (1987) and Engel and West (2005). It is shown that the long memory induced by learning can account for puzzling features often encountered in empirical work, in the context of predictive regressions for asset returns and the forward premium anomaly. Proofs are given in the appendix at the end. Supplementary material collecting further proofs and simulation results is available online.

Throughout the paper, $f(x) \sim g(x)$ as $x \to a$ means $\lim_{x \to a} f(x)/g(x) = 1$. Also, we use the notation $\mathsf{sd}(X)$ to refer to the standard deviation $\sqrt{var(X)}$.

2 Framework

2.1 Model and assumptions

In this paper, we consider a simple forward-looking model:

$$y_t = \beta y_{t+1}^e + x_t, \quad t = 1, 2, ..., T \tag{1}$$

where y_{t+1}^e denotes the expectation of y_{t+1} conditional on information up to time t, and x_t is an exogenous process that drives y_t. Under rational expectations, $y_{t+1}^e = E_t(y_{t+1})$, where E_t denotes expectations based on the true law of motion of y_t.

It is well-known that when $|\beta| < 1$ and $\lim_{T\to\infty} E_t(y_T) < \infty$, the rational expectations equilibrium (REE) satisfies

$$y_t = \sum_{j=0}^{\infty} \beta^j E_t(x_{t+j}), \tag{2}$$

provided this sum converges, which depends on the properties of x_t. We consider the leading case where there are no exogenous dynamics, i.e., where $E_t(x_{t+j}) = \mu$, for some constant μ and for all $j > 0$. The motivation for excluding exogenous dynamics is to focus on the persistence induced endogenously through learning. In our analytical results we will consider a more general exogenous process x_t, see assumption B below.

Let $\epsilon_t = x_t - E_{t-1}(x_t)$. Then, from equation (1), the REE can be expressed as:

$$y_t = \alpha + \epsilon_t, \tag{3}$$

where $\alpha = \mu/(1-\beta)$. Under rational expectations, $y_{t+1}^e = \alpha$. However, if agents do not know the mean of y_t, α, they must use past data to learn about it, so y_{t+1}^e will depend on the data they use.

We consider a representative agent who forms her expectations y_{t+1}^e using the available sample $y_1, ..., y_t$ using a linear algorithm of the form:

$$y_{t+1}^e = \sum_{j=0}^{t-1} \kappa_{t,j} y_{t-j} + \zeta_t. \tag{4}$$

where the term ζ_t represents the impact of the initial beliefs. Our main motivation for focusing our attention on linear learning algorithms is to emphasize that long range dependence can

arise without the need for nonlinearities – contrast this with Davidson and Sibbertsen (2005) and Miller and Park (2010) (see also the surveys by Granger and Ding, 1996, and Davidson and Teräsvirta, 2002). We use a representative agent framework to avoid inducing long memory through heterogeneity and aggregation, as in e.g. Granger (1980), Abadir and Talmain (2002) and Zaffaroni (2004). We define the polynomial $\kappa_t(L) = \sum_{j=0}^{t-1} \kappa_{t,j} L^j$ where L is the lag operator. To quantify how much the agent discounts past observations when forming her expectations, we use the mean lag of κ_t, which is defined as $m(\kappa_t) = [\kappa_t(1)]^{-1} \sum_{j=1}^{t-1} j\kappa_{t,j}$. We make the following assumptions about the learning algorithm:

Assumption A.

A.1. $(\kappa_{t,j})$ is nonstochastic;

A.2. $\kappa_t(1) \leq 1$ for all t and as $t \to \infty$;

A.3. There exists $m_\kappa > 0$ and $\delta_\kappa \in [0,1]$ such that $m(\kappa_t) \sim m_\kappa t^{\delta_\kappa}$, as $t \to \infty$.

Assumption A.1 precludes cases in which agents use additional regressors in the forecasting model, such as when they estimate an autoregressive model. It is made for simplicity of the analysis and restricts the scope of the paper, but we show below that it still allows for a wide range of models. Assumption A.2 is a common feature of most learning algorithms. It implies in particular that $\kappa_{t,t-1} \to 0$ as $t \to \infty$. Under assumption A.3 $\lim_{t\to\infty} \frac{\log m(\kappa_t)}{\log t}$ exists. This precludes cases where there exists a slowly varying function S_κ (i.e. where $\lim_{t\to\infty} S_\kappa(\lambda t)/S_\kappa(t) = 1$ for $\lambda > 0$) such that $m(\kappa_t) \sim m_\kappa t^{\delta_\kappa} S_\kappa(t)$. This is inconsequential to our analysis (although it will exclude some parameter values in section 3 but simplifies the exposition since $\delta_\kappa = 0$ here implies that $m(\kappa_t)$ is bounded.

We quantify the magnitude of decay of the weights $(\kappa_{t,j})$ via the parameter δ_κ as in the following definition:

Definition. *The parameter $\delta_\kappa \in [0,1]$ defined in assumption A.3 is referred to as the length of the learning window. The learning window is said to be short when $\delta_\kappa = 0$ and long otherwise.*

The distinction between short window (SW) and long window (LW) learning plays a significant role in our analytical results.

2.2 Examples

In the following examples of learning algorithms, we show that both short and long window learning may arise in standard settings of discounted least-squares, and give a further motivation via a model of beliefs. The analytical results we derive in the next section hence apply to the prototypical algorithms used in the literature.

2.2.1 Discounted Least-Squares

Both SW and LW learning may arise in the context of weighted, or discounted, least-squares (DLS), see Sargent (1999), where agents solve

$$y_{t+1}^{e} = \underset{\tau}{\operatorname{argmin}} \sum_{j=0}^{t-1} w_{t,j} \left(y_{t-j} - \tau\right)^2 \tag{5}$$

thus yielding $\kappa_{t,j} = \left(\sum_{i=0}^{t-1} w_{t,i}\right)^{-1} w_{t,j}$ and $\zeta_t = 0$ in the learning algorithm (4). Algorithms which arise as discounted least squares share the property that Sargent (1993, p.19) writes as the "sum of the weights equals unity". This corresponds here to $\kappa_t(1) = 1$, and satisfies the summability condition of assumption A.2.

The class of SW algorithms comprises all DLS algorithms with weights that decay fast enough, i.e. where $w_{t,j} = o\left(j^{-2}\right)$ as $j \to \infty$, e.g. exponentially decaying weights (such as with exponential smoothing), sufficiently quickly decaying hyperbolic weights or any fixed-window estimator (such as the rolling windows considered in Giacomini and White, 2006).

LW algorithms are such that $j^2 w_{t,j} \to \infty$ as $t, j \to \infty$ with $j \leq t$.[1] Recursive least-squares corresponds to no discounting of past observations, $w_{t,j} = 1$, and so belongs to this class, with $\delta_\kappa = 1$. The algorithms where only a fraction of the sample size t is used, such as when $w_{t,j} = 1$ for $j \leq t^\nu$, $\nu \in (0,1)$ are also LW learning algorithms with $\delta_\kappa = \nu$.

2.2.2 A simple model of beliefs

In the terminology of the learning literature, the model on which agents base their forecasts is referred to as the perceived law of motion (PLM). In order to understand how short or

[1] The specific case where $j^2 w_{t,j}$ remains bounded and nonzero yields a mean lag $m(\kappa_t) = O(\log t)$ which does not satisfy assumption A.3. We therefore exclude this from our analysis.

long window learning might arise from agents' beliefs and information set, consider a PLM based on a slightly generalized version of (3):

$$y_t = \alpha_t + \epsilon_t, \tag{6a}$$

$$\alpha_t = \alpha_{t-1} + q_t v_t, \quad t \geq 1, \tag{6b}$$

where q_t is a binary random variable with $\Pr(q_t = 1) = p$, and ϵ_t and v_t are independent and *i.i.d.*, with mean zero and variance normalized to 1. This PLM is known as a 'mean-plus-noise' model and allows the mean of the process y_t to change over time. When $p = 0$, agents assume the mean to be constant, i.e., $\alpha_t = \alpha_0$ for all t. This PLM clearly nests the rational expectations equilibrium (3).

Under the PLM (6), the optimal estimate a_t of α_t as a linear function of current and past data on y_t is given by the Kalman Filter (see Durbin and Koopman, 2008), which here takes the form:[2]

$$a_t = a_{t-1} + g_t (y_t - a_{t-1}), \quad t = 1, 2, ..., T, \tag{7a}$$

$$g_t = \frac{g_{t-1} + p}{1 + g_{t-1} + p}, \quad t \geq 2, \;\; g_1 = \frac{\sigma_0^2}{1 + \sigma_0^2} \tag{7b}$$

with σ_0^2 measuring the variance of agents' prior beliefs about α. The parameter σ_0^2 can also be interpreted as inversely related to agents' confidence in their prior expectation of α, given by a_0. g_t is the so-called gain sequence. The learning algorithm can be rewritten as

$$y_{t+1}^e = a_t = a_0 \prod_{i=0}^{t-1} (1 - g_{t-i}) + \sum_{j=0}^{t-1} \left[g_{t-j} \prod_{i=0}^{j-1} (1 - g_{t-i}) \right] y_{t-j} \tag{8}$$

which takes the form (4) with $\zeta_t = a_0 \prod_{i=0}^{t-1} (1 - g_{t-i})$ and $\kappa_{t,j} = g_{t-j} \prod_{i=0}^{j-1} (1 - g_{t-i})$.

Expression (7b) shows that discounting of past observations increases with the perceived probability of breaks. When $p = 0$, $g_t \to 0$ for all σ_0^2, and this is referred to as decreasing gain learning. Recursive Least Squares (RLS), i.e. $g_t = 1/t$ and $a_t = t^{-1} \sum_{j=1}^{t} y_t$ – alternatively $\zeta_t = 0$ and $\kappa_{t,j} = 1/t$ in (4) – is a special case when the prior is diffuse ($\sigma_0^2 = \infty$). With RLS, a_0 has no effect on the updating algorithm – intuitively, initial beliefs are so imprecise that they are not taken into account in subsequent inference. $\sigma_0^2 < \infty$ can also be interpreted as reflecting information in some prior sample. For example, setting $\sigma_0^2 = 1/t_0$, where t_0 is

[2] In the interest of simplicity, we do not consider nonlinear filtering, which would be more efficient.

an integer, makes $g_t = (t + t_0)^{-1}$. Assuming that the mean of y_t is constant therefore leads agents to use a LW learning algorithm with $\delta_\kappa = 1$.

When $p > 0$, i.e. when agents perceive that there is a nonzero probability that the mean might change, the learning algorithm has short window. To quote Giacomini and White (2006) "when there is inadequately modeled heterogeneity, observations from the more distant past may lose their predictive relevance. Alternatively, when dynamics are inadequately modeled, a limited-memory estimator can better track a series of interest". If agents suspect that the mean of y_t may be nonconstant and shift in an unanticipated manner, it may be preferable for them to use a short-window algorithm, thus achieving some robustness to dynamic misspecification. The gain parameter g_t converges to a constant $\bar{g} = \left(\sqrt{p(p+4)} - p\right)/2 > 0$ for all σ_0^2. Constant gain learning arises when σ_0^2 is chosen such that $g_1 = \bar{g}$. When $g_1 \neq \bar{g}$, the gain converges exponentially fast to its limit.[3] This algorithm is also referred to as perpetual learning, and it is quite popular in the empirical literature, see, e.g., Chakraborty and Evans (2009). This learning algorithm is close to (equal to, when $g_1 = \bar{g}$) the adaptive expectations framework of Cagan (1956) and Nerlove (1958) and the exponential smoother class of Muth (1960), see also Cogley (2002).

Another type of long window learning was suggested and empirically evaluated by Malmendier and Nagel (2011) where agents form their expectations discounting past observations with time-varying gain $g_t = \frac{\theta}{t}$ with $\theta > 1$. This algorithm implies that for both $t, t-j \to \infty$, the weights decay hyperbolically with $j : \kappa_{t,j} \sim \frac{\theta}{(t-\theta)^\theta} \frac{(t-j-\theta)^\theta}{(t-j)}$ and the algorithm is of long window type with $\delta_\kappa = 1$.

The dynamics of y_t under learning are determined by the actual law of motion (ALM). In the present example, the ALM is given by:

$$y_t = \frac{1 - g_t}{1 - \beta g_t} \beta a_{t-1} + \frac{x_t}{1 - \beta g_t}, \quad \text{for } t = 1, 2, ..., T. \tag{9}$$

[3]The gain sequence follows the recursion $g_t = G_p(g_{t-1})$ for $t > 1$ where G_p is homographic. The fixed points $G_p(g) = g$ are $\left(\pm\sqrt{p(p+4)} - p\right)/2$ when $p > 0$ and zero if $p = 0$. For $p > 0$, denote $\bar{g}, \breve{g}$ the positive and negative solutions. Then letting $\phi_t = \frac{g_t - \bar{g}}{g_t - \breve{g}}$, it follows that $\phi_t = \varrho_p^{t-1}\phi_1$, $\varrho_p = \frac{\breve{g}+1+p}{\bar{g}+p}$ and g_t converges exponentially fast to its positive limit:

$$g_t - \bar{g} = \sqrt{p(p+4)} \frac{\varrho_p^{t-1}\phi_1}{1 - \varrho_p^{t-1}\phi_1}.$$

The analytical results we provide below show how agents' perceptions and hence their choice of algorithm defining y_{t+1}^e affect the dynamics of y_t, and in particular its low frequency variability.

2.3 Exogenous dynamics

So far, we have assumed that the process x_t is not predictable, so there are no exogenous dynamics introduced through x_t. This was sufficient to motivate learning algorithms of the form (4) as ways to learn about the parameters of the forecasting model that would arise under rational expectations: with unpredictability of x_t, the rational forecast of y_{t+1} would be its unconditional mean. However, it is relatively straightforward to allow for some persistence in x_t while we still maintain the learning algorithm (4). So, we make the following assumption about x_t, which is less restrictive than the one we used to motivate the learning algorithm.

Assumption B. The process x_t is covariance stationary with finite fourth moments. Its spectral density is differentiable everywhere; it is nonzero and flat at the origin. Its autocovariance function decays exponentially.

Assumption B characterizes a typical covariance stationary process with short memory. It is clearly satisfied by processes that admit a finite-order invertible autoregressive moving average (ARMA) representation. This restriction ensures that long-range dependence is not introduced exogenously into the model, yet it allows for some dependence in x_t.

We need to point out that if x_t is persistent, and therefore predictable, then lags of x_t (or lags of y_t if x_t is unobserved) become useful in forecasting y_t. Therefore, with exogenous dynamics, the rational expectations equilibrium is not given by (3), and the mean-plus-noise model introduced in the previous subsection does not nest it. So, the learning algorithms (4) cannot be thought of as providing information about the rational expectations equilibrium. Such cases can still admit an interesting interpretation in terms of the notion of a restricted perceptions equilibrium (RPE, see Sargent, 1993). Extension of our analysis to more general learning algorithms that nest REEs with exogenous dynamics is both difficult from an analytical point of view, and without a strong empirical motivation. Indeed in many applications, e.g. in section 5, the assumption that x_t is unpredictable is plausible, and x_t has often been considered so by past authors. Another motivation for simplicity is that the

aforementioned learning algorithms are necessarily nonlinear, and therefore may confound the effect of learning with that of nonlinearity when we look at the low frequency properties of the data.

3 Analytical results

This section provides our main results. We analyze the impact of the learning window length on the memory of the resulting process. For clarity, we start with learning algorithms whose coefficients are time-invariant, i.e., $\kappa_{t,j} = \kappa_j$ for all t in (4). These algorithms have the property that the learning window has length $\delta_\kappa < 1$. We then analyze the case $\delta_\kappa = 1$ in which assumption A implies that the coefficients of the learning algorithm must be time-varying and we focus on RLS. Finally, we consider the case of perpetual learning in the empirically relevant case where the gain is very small.

3.1 Long memory

We start by providing our working definition of long memory. There are several measures of dependence that can be used to characterize the memory of a stochastic process, such as mixing coefficients and autocorrelations (when they exist). Various alternative definitions of short memory are available (e.g., various mixing conditions, see White, 2000). These definitions are not equivalent, but they typically imply that short memory requires that the variance of partial sums, scaled by the sample size, T, should be bounded.[4] If this does not hold, we will say that the process exhibits long memory. This is the definition adopted by Diebold and Inoue (2001) in their study of the connection between structural change and long memory. Analogously to our previous discussion of the length of the learning window, we can also define the 'degree of memory' of a process z_t by the smallest d (when it exists) such that

$$\mathsf{sd}\left(T^{-1/2}S_T\right) = O\left(T^d\right), \quad \text{where } S_T = \sum_{t=1}^{T} z_t. \tag{10}$$

[4] Any definition of short memory that implies an invariance principle satisfies the restriction on the variance of partial sums, e.g., Andrews and Pollard (1994), Rosenblatt (1956), or White (2000).

If $d = 0$, the process exhibits short memory, while $d > 0$ corresponds to long memory ($d < 0$ is sometimes referred to as antipersistence).[5]

The above definition applies generally to any stochastic processes that have finite second moments (which we assume in this paper). For a covariance stationary process, where the autocorrelation function is a common measure of persistence, short memory requires absolute summability of its autocorrelation function, or a finite spectral density at zero. Thus, long memory arises when the autocorrelation coefficients are non-summable, or the spectrum has a pole at frequency zero. This gives rise to the following definitions of d, that are equivalent to (10) for covariance stationary processes, see Beran (1994) or Baillie (1996):

$$\begin{aligned} &\rho_z(k) \sim c_\rho k^{2d-1}, \qquad \text{as } k \to \infty \\ &f_z(\omega) \sim c_f |\omega|^{-2d}, \quad \text{as } \omega \to 0, \end{aligned} \tag{11}$$

for some positive constants c_ρ, c_f, where $\rho_z(k) = \mathsf{Corr}[z_t, z_{t+k}]$ is the autocorrelation function (ACF) of a covariance stationary stochastic process z_t and $f_z(\omega)$ is its spectral density. For $d > 0$, the autocorrelation function at long lags and the spectrum at low frequencies have the familiar hyperbolic shape that has traditionally been used to define long memory.

Fractional integration, denoted I(d), is a well-known example of a class of processes that exhibit long memory. When $d < 1$, the process is mean reverting (in the sense of Campbell and Mankiw, 1987, that the impulse response function to fundamental innovations converges to zero, see Cheung and Lai, 1993). Moreover, I(d) processes admit a covariance stationary representation when $d \in (-1/2, 1/2)$, and are non-stationary if $d \geq 1/2$. Long range dependence, or long memory, arises when the degree of fractional integration is positive, $d > 0$. In the case of nonstationary processes, the ACF definition of d in (11) does not apply,[6] so we use the ACF/spectrum of Δz, as in Heyde and Yang (1997):

$$\begin{aligned} &\rho_{\Delta z}(k) \sim c_\rho k^{2(d-1)-1}, \quad 1/2 < d < 1 \qquad \text{as } k \to \infty; \\ &f_{\Delta z}(\omega) \sim c_f |\omega|^{-2(d-1)}, \quad 1/2 < d < 1 \qquad \text{as } \omega \to 0. \end{aligned} \tag{12}$$

[5] In the context of nonlinear cointegration, Gonzalo and Pitarakis (2006) have introduced the terminology "summable of order d" for processes that satisfy the definition given in equation (10) above, see also Berenguer-Rio and Gonzalo (2011).

[6] The property $f_z(\omega) \sim c_f |\omega|^{-2d}$ can be applied also to nonstationary cases with $1/2 < d < 1$ if $f_z(\omega)$ is defined in the sense of Solo (1992) as the limit of the expectation of the sample periodogram.

Müller and Watson (2008) study the low frequency properties of economic time series, and show that there exist many statistical models which may be used to model strong persistence short of a stochastic trend. These constitute, in the words of the authors "continuous bridges between the I(0) and I(1) models". Müller and Watson suggest that a good measure of the distance of models to the I(0) and I(1) boundaries is the total variation distance between the implied measures.[7] This could be used as an alternative way to define and measure the degree of memory of a stochastic process. Unfortunately, this does not seem to be analytically tractable in the models that we study in this paper.

Estimators of d exist both in the time and frequency domains. In the time domain, estimation has historically been performed via the "rescaled range-statistic" R/S of Hurst (1951) (see also Baillie, 1996, and Lo, 1991) or, in the context of non-stationary process via the decay long-run variance, see e.g. Teverovsky and Taqqu (1997) in the context of shifting means and declining trends. Yet, it is more common to estimate d via the shape of the spectral density close to the origin. Two classes of estimators have been proposed in the literature, either maximizing the local "Whittle" likelihood (see Robinson, 1995) or by regressing an estimate of the log spectral density on the log of (functions of) the frequency, see Geweke and Porter-Hudak (1983) (GPH henceforth) and Robinson (1995b). Denoting by $\widehat{f}(\omega_j)$ an estimator of the spectral density[8] evaluated at the jth Fourier frequency $\omega_j = 2\pi j/T$, the latter consists of estimating the regression:[9]

$$\log \widehat{f}(\omega_j) = \varphi_0 + \varphi_1 \log \omega_j + \varepsilon_{\omega,j}, \quad j = 1, ..., n \tag{13}$$

where n is a truncation parameter, which must be chosen such that $n/T \to 0$. The estimator of the fractional integration parameter d is given by $-\widehat{\varphi}_1/2$, where $\widehat{\varphi}_1$ is the OLS estimate of φ_1. Alternative estimators can be obtained based on different choices of truncation parameter n, different regressors or (smoothed) estimators of $\widehat{f}(\omega_j)$.

In the following, we use all of the above three characterization of long memory, where possible, for the processes under consideration.

[7] The total variation distance between two probability measures is defined as the largest absolute difference the two probability measures assign to the same event, maximized over all events.

[8] The literature has considered the sample periodogram or smoothed estimates thereof.

[9] The recent literature (e.g. Hurvich et al. (1998), Kim and Phillips (1999) and Phillips (2007) advocates the use of $\log\left|1 - e^{i\omega_j}\right|$ as a regressor instead of $\log \omega_j$ but this choice does not affect the results of the paper.

3.2 Degree of memory under Rational Expectations

Under rational expectations, if x_t satisfies assumption B, then so does y_t when $\beta \leq 1$, see e.g. Gourieroux et al. (1982). We can generalize this to show the conditions for the dependence of x_t such that y_t does not exhibit long memory.

Proposition 1 *Let x_t admit an infinite moving average representation: $x_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}$, where ϵ_t is i.i.d with zero expectation and finite variance. Then the solution to $y_t = \beta E_t y_{t+1} + x_t$ with $\beta \leq 1$ satisfies*

$$\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right) = O(1),$$

if $|\beta| < 1$ and $\left|\sum_{j=0}^{\infty}\theta_j\right| < \infty$, or if $\beta = 1$ and $\sum_{j=0}^{\infty} j\,|\theta_j| < \infty$.

In the proof of the above result, we find that long memory in y_t can arise in the following case: $\beta = 1$ and $\{\theta_j\}$ satisfy $\left|\sum_{j=0}^{\infty}\theta_j\right| < \infty$, $\sum_{j=0}^{\infty} j\,|\theta_j| = \infty$ and $\sum_{j=0}^{\infty}\theta_j \neq 0$. In this situation, x_t exhibits short memory by our definitions, but its behavior in finite samples is very similar to that of a long memory process. Examples of such processes are rare in the literature and they are often assumed away (see e.g. Phillips and Magdalinos, 2005, or Perron and Qu, 2007) as their spectral density is nonzero yet not differentiable at the origin (see Stock, 1994) so that they are in finite sample difficult to distinguish from long memory process. Thus, with the exception of such pathological situations, long memory cannot arise endogenously under rational expectations.

3.3 Learning algorithms with constant coefficients

Consider model (1), and assume that the linear learning algorithm (4) has constant coefficients $\kappa_{j,t} = \kappa_j$ for all $t \geq 0$. Assumption A.2 implies that $\kappa_j = o\left(j^{-1}\right)$, and the length of the learning window δ_κ depends on the rate of decay of the weights. If $\kappa_j = O\left(j^{-2}\right)$, the learning window is short under assumption A.3, while if $\kappa_j \sim c_\kappa j^{\delta_\kappa - 2}$, for some $c_\kappa > 0$ and $0 < \delta_\kappa < 1$, the learning window is long, with length δ_κ.

For simplicity, we assume that there is an infinite history of $\{y_t\}$ and define $y_t^- = y_t 1_{\{t\leq 0\}}$ and $y_t^+ = y_t - y_t^-$. To ensure existence of the initial beliefs, we define ζ_t such that $\zeta_t = \kappa(L)\, y_t^- = \sum_{j=t}^{\infty}\kappa_j y_{t-j}$ if $\delta_\kappa \in (1/2, 1)$ and $\Delta\zeta_t = \kappa(L)\,\Delta y_t^-$ if $\delta_\kappa \in (0, 1/2)$. Owing to

non-stationarity, we will be led later to restrict further the assumption on ζ_t when letting $\beta \to 1$. A simplifying assumption often made in the literature is $y_t^- \equiv 0$, see e.g. Diebold and Rudebusch (1991) and Tanaka (1999). Yet, it has been shown that this assumption (which is related to the difference between Type I and Type II Fractional Brownian motions) is not innocuous for the definition of the spectral density, so we avoid it: see Marinucci and Robinson (1999), Davidson and Hashimzade (2008, 2009).

Under the previous assumptions, the ALM can be written, for $t \leq T$, as

$$\begin{aligned} &(1 - \beta\kappa(L))\, y_t = x_t, \qquad \text{if } \delta_\kappa \in (1/2, 1) \\ &(1 - \beta\kappa(L))\, \Delta y_t = \Delta x_t \quad \text{if } \delta_\kappa \in (0, 1/2) \end{aligned} \tag{14}$$

It is also clear in the previous expressions that $(1 - \beta\kappa(1))\, E(y_t) = E(x_t)$ so (14) can be expressed in deviation from the expectations. In other words, we may assume without loss of generality and for ease of exposition that $E(x_t) = 0$.

3.3.1 Long Memory under 'local-to-unity' asymptotics

Under the assumption that x_t is a short memory process, the persistence of y_t in terms of low frequency variability depends on the learning algorithm $\kappa(\cdot)$ and on the coefficient β. When $\beta\kappa(1) = 1$, the autoregressive lag polynomial $(1 - \beta\kappa(L))$ has a unit root. Because in typical applications β is interpreted as a discount factor that is close to 1, and the learning algorithm is least squares yielding $\kappa(1) = 1$, the relevant framework of analysis is a local-asymptotic nesting in which β is modelled as local to unity, in the sense that $1 - \beta_T = O(T^{-\nu})$, where T is the sample size and $\nu > 0$. Formally, this means that the stochastic process of y is a triangular array $\{y_{t,T}\}_{t \leq T}$. However, we shall omit the dependence of β and y_t on T for notational simplicity. For ease of exposition, and without loss of generality, we assume in the following that $\kappa(1) = 1$.

The motivation for working under the local-to-unity asymptotic framework is twofold. First, it provides a better description of the behavior of the process y_t for many values of β when T is finite. Second, it characterizes the class of situations in which the process is virtually indistinguishable from a process with a unit root. For completeness, we give in a supplementary appendix some analytical results on the behavior of the spectrum near the zero frequency when $\beta < 1$ and fixed.

Local-to-unity asymptotics or triangular-array asymptotics are common in the time series literature. To make the connection with this literature, consider the autoregressive process of order 1, AR(1). This is a special case of (14) with $\kappa(L) \equiv L$ and x_t white noise, so that β is the autoregressive coefficient. When $\beta \in (-1, 1)$, y_t is covariance stationary with short memory (i.e. exponentially decaying autocovariances), while $\beta = 1$ corresponds to the usual random walk setting. It is well known that when β is close to – though strictly below – unity, asymptotic approximations computed under the covariance stationarity assumption are less accurate in finite samples than asymptotics based on the unit-root setting. To solve this issue, a number of authors proposed to express the proximity of β to unity in terms of the available sample size, letting $c_\beta \geq 0$ such that $\beta_T = 1 - c_\beta/T$, see Bobkoski (1983), Cavanagh (1985), Chan and Wei (1987) and Phillips (1987). These articles spurred a large literature in which some authors suggested that the previous assumption could be extended to

$$1 - \beta_T = c_\beta T^{-\nu}, \quad \text{for } \nu > 0. \tag{15}$$

Giraitis and Phillips (2006) and Phillips and Magdalinos (2007) showed that when $\nu < 1$, the autoregressive root is different enough from unity to ensure that distributions are similar to the covariance stationary setting, $\nu = 0$, (e.g. asymptotic normality as opposed to Dickey-Fuller type distributions; see also Phillips, Magdalinos and Giraitis, 2010). $\nu \in (0, 1)$ is referred to as the near-stationarity region. By contrast, Andrews and Guggenberger (2007) showed that when $\nu > 1$, (the "very nearly unity" region) the rates of convergence and distributions are comparable to the exact unit root case $\nu = +\infty$. To summarize, in the AR(1) model, when β is within a T^{-1} neighborhood of unity, the process y_t is close to an $I(1)$ process and when β lies strictly outside any such neighborhood, asymptotic distributions (such as those of the estimators of autoregressive coefficients and associated t-statistics) resemble those of covariance stationary processes.

The following result gives the memory properties of the process y_t in terms of the order of magnitude (as T increases) of the long run variance of y_t, in accordance with the definition (10) of the degree of memory.

Theorem 2 *Consider the model* $y_t = \beta y_{t+1}^e + x_t$, *with* $y_{t+1}^e = \kappa(L)\, y_t^+ + \zeta_t$ *and* $\sum_{t=1}^T \zeta_t = o_p\left(\sum_{t=1}^T x_t\right)$. *Suppose* x_t *satisfies assumption B, with* $E(x_t) = 0$, *the learning algorithm*

$\kappa(\cdot)$ satisfies assumption A, with $\delta_\kappa \in [0,1)$, $\delta_\kappa \neq 1/2$, $\kappa(1) = 1$, and $\beta = 1 - c_\beta T^{-\nu}$ with $\nu \in [0,1]$ and $c_\beta > 0$. Then, as $T \to \infty$, $S_T = \sum_{t=1}^{T} y_t$ satisfies:

$$\mathsf{sd}\left(T^{-1/2} S_T\right) = O\left(T^{\min(\nu, 1-\delta_\kappa)}\right).$$

The above result shows the memory of the process y_t depends on (i) the proximity of β to unity and (ii) the length of the learning window. If $\nu = 0$, the process exhibits short memory, irrespective of the learning window.[10] For $\nu > 0$, the memory of the process depends on whether $\nu \leq 1 - \delta_\kappa$ or $\nu > 1 - \delta_\kappa$, i.e., on how close β is to unity relative to the length of the learning window.

When β is sufficiently close to unity, $\nu > 1 - \delta_\kappa$, we can derive expressions for the spectral density of y_t at low frequencies and the rate of decay of its autocorrelation function that accord with the alternative definitions of long memory given in equation (11).

Theorem 3 *Under the assumptions of theorem 2, and if $\nu > 1 - \delta_\kappa$, then:*
1. the spectral density f_y of y_t evaluated at Fourier frequencies $\omega_j = 2\pi j/T$ with $j = 0, ...n$, and $n = o(T)$, satisfies as $T \to \infty$,

$$f_y(\omega_j) \sim f_x(0)\, \omega_j^{-2(1-\delta_\kappa)}$$

2. the autocorrelation functions ρ_y of y_t , or $\rho_{\Delta y}$ of Δy_t, evaluated at $k = o(T)$, satisfy as $T, k \to \infty$,

$$\begin{aligned} \rho_y(k) &= O\left(k^{1-2\delta_\kappa}\right) \qquad \text{if } \tfrac{1}{2} < \delta_\kappa < 1 \\ \rho_{\Delta y}(k) &= O\left(k^{-2\delta_\kappa - 1}\right) \quad \text{if } 0 < \delta_\kappa < \tfrac{1}{2}. \end{aligned}$$

Both theorems show that the persistence of the process y_t is a function of the relative values of the length of the learning window and the proximity of β to unity. When β is sufficiently close to unity, the memory of the process is determined entirely by the length of the learning window, δ_κ, and is inversely related to δ_κ.[11]

[10] In the case $\nu = 0$, it can be shown that the derivative of the spectrum of y_t at zero is unbounded, as in fractionally integrated processes, and that the value of the spectrum is positive and increasing in β, so in finite samples it may be difficult to distinguish y_t from a fractionally integrated process. These results, which are avaiable in the supplementary material to this paper, provide additional motivation for focusing on the local-to-unity case.

[11] This arises because the window length δ_κ is positively related to how much weight agents put on distant

3.3.2 Consistency of the estimator of the degree of memory

Having established the long memory implications of learning dynamics, we now turn to the properties of the GPH estimator of the long memory parameter d. We rely on the high level assumption that there exists a spectral density estimator that is consistent at low frequencies. Sufficient conditions for this assumption for long memory processes can be found at various places in the literature, see e.g. Robinson (1994b), and specifically for $\delta_\kappa \in (1/2, 1)$, Robinson(1994a) and Delgado and Robinson (1996). The following result establishes conditions under which this estimator is consistent for the value implied by the length of the window of the learning algorithm δ_κ.

Theorem 4 *Under the model and assumptions of theorem 2 with* $\nu > 1 - \delta_\kappa$, *let* $\widehat{f}_{y,T}$ *and* $\widehat{f}_{\Delta y,T}$ *denote estimators of the spectral densities* f_y *and* $f_{\Delta y}$. *Let* $n = o(T)$ *and assume that for all Fourier frequencies* ω_j, $j = 1, ..., n$:
if $\delta_\kappa \in (1/2, 1)$, $plim_{T\to\infty} \widehat{f}_{y,T}(\omega_j) / f_y(\omega_j) \to 1$, *or*
if $\delta_\kappa \in (0, 1/2)$, $plim_{T\to\infty} \widehat{f}_{\Delta y,T}(\omega_j) / f_{\Delta y}(\omega_j) \to 1$.
Consider regressing $\log \widehat{f}_{y,T}(\omega_j)$, *if* $\delta_\kappa \in (1/2, 1)$, *or* $\log \widehat{f}_{\Delta y,T}$, *if* $\delta_\kappa \in (0, 1/2)$, *on a constant and* $-2 \log \omega_j$ *over the ordinates* $j = 1, ..., n$. *Then the estimator* $\widehat{d}$ *of the coefficient of* $-2 \log \omega_j$ *in the regression satisfies as* $n \to \infty$,

$$\widehat{d} \xrightarrow{p} \begin{cases} 1 - \delta_\kappa & \text{if } \delta_\kappa \in (1/2, 1), \\ \delta_\kappa & \text{if } \delta_\kappa \in (0, 1/2). \end{cases}$$

An interesting implication of this theorem is that it supports the notion of a self-confirming or consistent expectations equilibrium, see Brock and Hommes (1997). If agents believed that the process y_t exhibited long memory and used a hyperbolically weighted moving average filter to generate their forecasts, such as $\kappa(L) = 1 - (1 - L)^{1-\delta_\kappa}$, the resulting dynamics of the data would indeed exhibit long memory, and agents' estimates of the fractional order would be consistent with their original beliefs.

observations. Yet, in fractionally integrated process, the degree of integration is a positive function of the speed of decay of the weights of the lag polynomial. Hence, the larger δ_κ the further away the resulting process y_t is a from a martingale (an I(1) process).

3.4 Recursive least squares

The results of the previous section concerned learning algorithms with constant coefficients and window length $\delta_\kappa < 1$. When $\delta_\kappa = 1$, summability of the coefficients of the learning algorithm means we must consider a learning algorithm with time-varying weights. It does not seem possible to provide general results for this case, but it is instructive to examine the important case of Recursive Least Squares (RLS).

In the mean-plus-noise model of section 2.2.2, RLS arises when agents' perceived probability of changes in the mean is zero, $p = 0$. When agents learn using RLS, the learning algorithm is nonconstant (specifically, $\kappa_{t,j} \sim 1/t$) and the resulting process is nonstationary. Under correct specification of the PLM by agents, learning converges to the REE and y_t itself tends to a weakly dependent process, see e.g., Evans and Honkapohja (2001). Yet the convergence can be so slow that y_t exhibits long memory as the following result shows.

Theorem 5 *Consider the model $y_t = \beta y_{t+1}^e + x_t$, where y_{t+1}^e is given by equation (8) with $g_t \sim 1/t$ and $a_0 = O_p(1)$, and suppose x_t satisfies assumption B. Then, as $T \to \infty$, $S_T = \sum_{t=1}^T y_t$ satisfies:*

$$\mathsf{sd}\left(T^{-1/2} S_T\right) = \begin{cases} O\left(T^{\beta-1/2}\right), & \text{if } 1/2 < \beta \leq 1, \\ O\left(\sqrt{\log T}\right), & \text{if } \beta = 1/2, \\ O(1), & \text{if } \beta < 1/2. \end{cases}$$

The theorem shows that the process exhibits long memory when $\beta > 1/2$. This explains a result from the learning literature on the properties of agents' forecasts under RLS learning: even though y_{t+1}^e converges to a constant when $\beta < 1$, asymptotic normality of y_{t+1}^e is only established when $\beta < 1/2$ (Evans and Honkapohja, 2001, theorem 7.10).

When $\beta = 1$, learning does not converge and persistence is strongest in that case. Unlike the previous results, long memory arises here without the use of local asymptotics. However, when β is close to 1, the behavior of the process in a finite sample may be better approximated by a local asymptotic framework. Because of the continuity as $\beta \to 1$ in theorem 5, a local-asymptotic setting would not modify the degree of memory of y_t.[12]

[12] A formal local-to-unity asymptotic approximation is available on the supplementary material to this paper. It is shown than for $\beta = 1 - c_\beta T^{-\nu}$, with $c_\beta > 0$ and $\nu \in (0, 1]$ the memory length is $d = 1/2$.

3.5 Perpetual learning with small gain parameter

Another leading example of a learning algorithm that features prominently in the empirical literature is CGLS, or perpetual learning, with a very small gain parameter. Often this type of learning induces behavior that is in some sense close to a rational expectations equilibrium, and it is sometimes referred to as 'nearly rational expectations', see Milani (2007). For fixed gain, CGLS is clearly a SW algorithm, but this is not an appropriate characterization when the gain parameter is small relative to the sample size. To make this precise, we consider a local-to-zero asymptotic nesting where the gain parameter goes to zero with the sample size. We focus for simplicity on the mean-plus-noise model of section 2.2.2 with $p > 0$. In that model, $y_{t+1}^e = a_t$, where a_t is an exponentially weighted moving average of past y_j, $j \leq t$. Specifically, under constant gain learning, a_t can be written as:

$$a_t = \left(1 - \frac{(1-\beta)\,\bar{g}}{1-\beta\bar{g}}\right)^t a_0 + \frac{\bar{g}}{1-\beta\bar{g}} \sum_{i=1}^{t} \left(1 - \frac{(1-\beta)\,\bar{g}}{1-\beta\bar{g}}\right)^{t-i} x_i. \tag{16}$$

So, if β is very close to unity or $\bar{g}$ very close to 0 such that $(1-(1-\beta)\,\bar{g}) \approx 1$, a_t exhibits near unit-root behavior. Yet, when $\bar{g}$ is small this attenuates the near-stochastic trend in a_t – as it appears before the summation in expression (16). We aim to characterize the implications for the dynamics of the processes of the proximity of $(\beta, \bar{g})$ to $(1, 0)$. We follow and extend the local-asymptotic approach of Chevillon *et al.* (2010) and let $1-\beta = c_\beta T^{-\nu}$ and $\bar{g} = c_g T^{-\lambda}$ for $(\nu, \lambda) \in [0,1]^2$. Larger values of ν and λ mean here that $1-\beta$ and $\bar{g}$ are assumed to lie in tighter neighborhoods of zero.

For $\lambda \leq 1$, the length of the learning window δ_κ is equal to λ, since the mean lag satisfies

$$m(\kappa_T) = O\left(T^\lambda\right), \tag{17}$$

see the proof in the appendix. Hence, $\bar{g} \to 0$ corresponds to long-window learning. The following theorem gives the implications for the memory of y_t.

Theorem 6 *Consider the model $y_t = \beta y_{t+1}^e + x_t$, with $y_{t+1}^e = a_t$ as given in equation (16), where x_t satisfies assumption B and $a_0 = O_p(1)$; the parameters $(\beta, \bar{g}) = \left(1 - c_\beta T^{-\nu}, c_g T^{-\lambda}\right)$, $(\lambda, \nu) \in [0,1]^2$ and c_β, c_g are positive. Then, as $T \to \infty$, $S_T = \sum_{t=1}^T y_t$ satisfies:*

$$\mathsf{sd}\left(T^{-1/2} S_T\right) = \begin{cases} O\left(T^{1-\lambda}\right), & \text{if } \nu > 1-\lambda; \\ O\left(T^\nu\right), & \text{if } \nu \leq 1-\lambda, \end{cases} \tag{18}$$

This result is entirely analogous to theorem 2, with $\lambda = \delta_\kappa$. Chevillon *et al.* (2010) studied only the case where $\nu = \lambda = 1/2$ and x_t is *i.i.d.* They did not consider the implications for the memory of y_t. Theorem 6 shows that perpetual learning with small gain create dynamics that are akin to long memory in finite samples.

4 Simulations

This section presents simulation evidence in support of the analytical results given above. We focus on the impact of the length of the learning window through the simple model of beliefs presented in section 2.2.2. We generate samples of $\{y_t\}$ from equation (9) for a relatively long sample of size $T = 1000$ and for various values of the parameters β and $\bar{g}$. We set $\sigma_0^2 = \infty$ (diffuse prior), so that a_0 is no longer relevant. We study the behavior of the variance of partial sums, the spectral density, and two popular estimators of the fractional differencing parameter d, the GPH and maximum local Whittle likelihood estimator.[13] We also report the power of tests of the null hypotheses $d = 0$ and $d = 1$. The exogenous variable x_t is assumed to be *i.i.d.* with mean zero and its variance is normalized to 1 without loss of generality. The number of Monte Carlo replications is 10,000. Additional figures reporting the rate of growth of the variance of partial sums and the densities of estimators of d are available in a supplementary appendix.

Figure 1 reports the Monte Carlo average log sample periodogram against the log frequency ($\log \omega$). This constitutes a standard visual evaluation of the presence of long range dependence if the log periodogram is linearly decreasing in $\log \omega$. When the learning algorithm is RLS, the figure indicates that y_t exhibits long memory for $\beta > 1/2$ and the degree of long memory increases with β. Table 1 records the means of the estimators, and the empirical rejection frequency (power) of tests of the hypotheses $d = 0$ and $d = 1$ (the latter is based on a test of $d = 0$ for Δy_t) against the one-sided alternatives $d > 0$ and $d < 1$ respectively. In addition to the fact that $E(\widehat{d})$ increases with β in accordance with theorem 5, we also observe that $E(\widehat{d})$ remains below unity.

Figure 1 and table 2 report the corresponding statistics when $\bar{g} > 0$ and the learning algorithm (with a random initial condition) approaches CGLS. The behavior of $E(\widehat{d})$ as well

[13] The Whittle estimator is obtained by constrained maximization over the range $d \in (-1, 2)$.

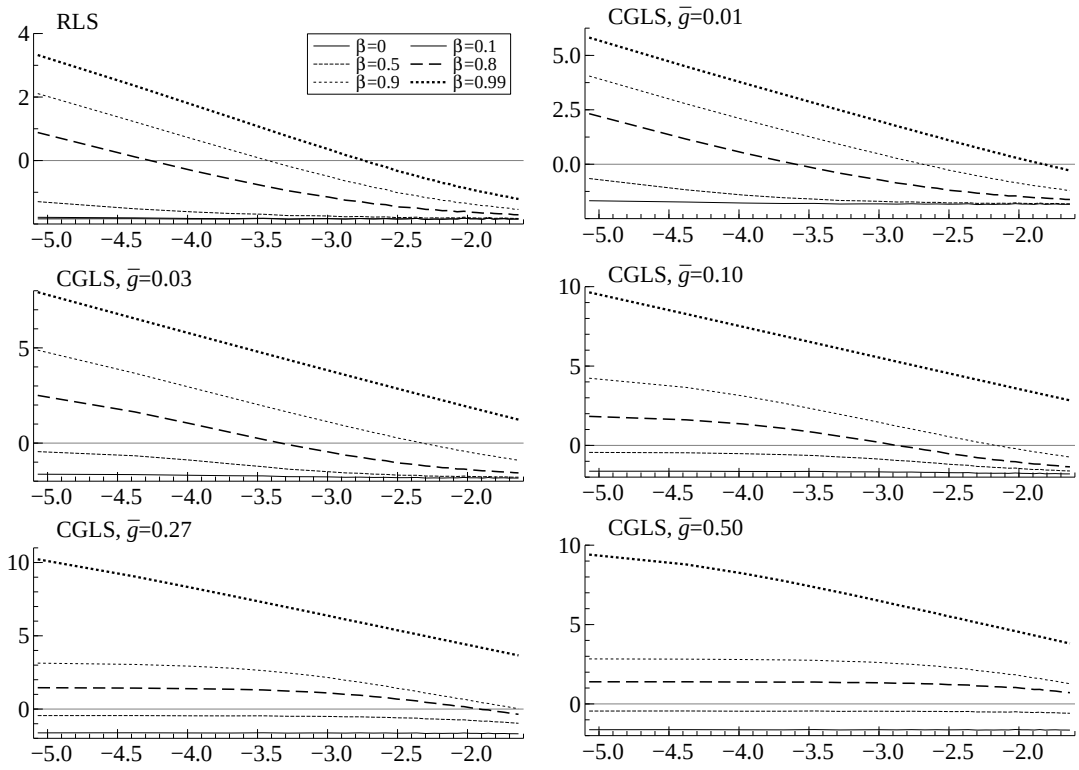

Figure 1: Monte Carlo averages of the log periodogram against the log of the first $\sqrt{T}$ Fourier frequencies with $T = 1,000$ observations. The model is the 'mean plus noise' perceived law of motion presented in section 2.2.2. The number of Monte Carlo replications is 10,000.

	Mean of $\widehat{d}$		Pr(Reject $d = 0$)		Pr(Reject $d = 1$)	
β	GPH	Whittle	GPH	Whittle	GPH	Whittle
0.00	0.001	-0.011	0.075	0.069	0.936	0.995
0.10	0.006	-0.007	0.081	0.077	0.922	0.993
0.50	0.055	0.039	0.179	0.182	0.795	0.950
0.80	0.291	0.245	0.656	0.677	0.561	0.753
0.90	0.438	0.378	0.805	0.817	0.466	0.635
0.99	0.573	0.510	0.890	0.899	0.374	0.519

Table 1: The table records the Monte Carlo mean of the estimator of the degree of fractional integration d and the empirical rejection frequency of the null hypotheses $\mathsf{H}_0 : d = 0$ and $\mathsf{H}_0 : d = 1$ for a nominal size of 5%. The tests use the Geweke & Porter-Hudak (1983) log periodogram regression estimator and the maximum local Whittle likelihood estimator of Robinson (1995a). The learning algorithm is RLS.

		Mean of $\widehat{d}$		Pr(Reject $d = 0$)		Pr(Reject $d = 1$)	
$\bar{g}$	β	GPH	Whittle	GPH	Whittle	GPH	Whittle
0.01	0.10	0.018	0.005	0.096	0.095	0.922	0.993
	0.50	0.119	0.104	0.319	0.364	0.795	0.950
	0.80	0.458	0.410	0.834	0.872	0.565	0.762
	0.90	0.657	0.599	0.930	0.948	0.474	0.655
	0.99	0.807	0.761	0.970	0.980	0.393	0.561
0.03	0.10	0.032	0.019	0.117	0.122	0.922	0.993
	0.50	0.194	0.181	0.525	0.626	0.794	0.946
	0.80	0.539	0.498	0.957	0.981	0.551	0.717
	0.90	0.770	0.720	0.990	0.996	0.449	0.597
	0.99	0.934	0.909	0.999	1.000	0.427	0.625
0.10	0.10	0.031	0.019	0.116	0.120	0.928	0.993
	0.50	0.216	0.212	0.598	0.717	0.821	0.955
	0.80	0.539	0.532	0.989	0.998	0.499	0.648
	0.90	0.765	0.741	1.000	1.000	0.296	0.406
	0.99	0.980	0.970	1.000	1.000	0.198	0.280
0.27	0.10	0.009	-0.003	0.085	0.083	0.939	0.995
	0.50	0.086	0.077	0.224	0.267	0.929	0.992
	0.80	0.322	0.330	0.843	0.939	0.731	0.879
	0.90	0.551	0.573	0.992	1.000	0.434	0.569
	0.99	0.971	0.967	1.000	1.000	0.029	0.045
0.50	0.10	0.003	-0.009	0.077	0.073	0.940	0.995
	0.50	0.021	0.009	0.100	0.102	0.955	0.997
	0.80	0.111	0.103	0.285	0.360	0.928	0.990
	0.90	0.268	0.272	0.733	0.861	0.792	0.924
	0.99	0.888	0.895	1.000	1.000	0.050	0.072

Table 2: The table records the Monte Carlo mean of the estimator of the degree of fractional integration d and the empirical rejection frequency of the null hypotheses $\mathsf{H}_0 : d = 0$ and $\mathsf{H}_0 : d = 1$ for a nominal size of 5%. The tests use the Geweke & Porter-Hudak (1983) log periodogram regression estimator and the maximum local Whittle likelihood estimator of Robinson (1995a). The learning algorithm is CGLS.

as $\Pr(\text{Reject } d=0)$ and $\Pr(\text{Reject } d=1)$ is non-monotonic in $\bar{g}$. As $\bar{g}$ increases, we observe first that persistence increases (nearly uniformly in β) up to $\bar{g}=0.10$ and then declines: this is in accordance with theorem 6. This is most evident from Table 2.

Unreported figures (available in the supplementary appendix) show that the log of $\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right)$ increases linearly with $\log T$ and that the growth rate of the ratio $\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right)/\log T$ tends quickly to the values the theorems imply for the degree of memory under both RLS learning and CGLS with local parameters. We also present there the densities of the estimators of d which complement the rejection probabilities recorded in tables 1 and 2.

5 Application to Present Value Models

We now consider the implications of learning in present value models. Specifically, we focus on the Campbell and Shiller (1987, 1988) models of stock prices and the term structure, and the model of Engel and West (2005) for exchange rates. Under rational expectations, both models are known to exhibit features that do not seem to match the data and have led to famous "puzzles". Many explanations for these puzzles have been proposed and some of them rely on some variables of interest presenting a large degree of persistence of exogenous origin. Here, we show that adaptive learning by agents may be generating the persistence necessary to explain the puzzling empirical features. Branch and Evans (2010), and Chakraborty and Evans (2008), also studied the potential of adaptive learning to explain those empirical puzzles. Our analysis differs from Chakraborty and Evans (2008), because we do not need to assume that fundamentals are strongly persistent, so our results are complementary to theirs. We also differ from Branch and Evans (2010) because they focus on explaining regime-switching in returns and their volatility, while we focus on their low frequency variation. Alternatively, Frankel and Froot (1987) use survey data to suggest an explanation to the forward premium anomaly via adaptive learning under heterogeneity.

The asset pricing models we consider admit the general formulation (see Engel and West, 2005):

$$Y_t = (1-\beta)\sum_{i=0}^{\infty}\beta^i E_t\left(\gamma_1' z_{t+i}\right) + \sum_{i=0}^{\infty}\beta^i E_t\left(\gamma_2' z_{t+i}\right), \tag{19}$$

where the log price Y_t is a function of the discounted sum of current and future vectors of 'fundamentals' z_t with impact coefficients γ_1 and γ_2.

Define $y_t = \Delta Y_t$ and the innovation $r_t = Y_t - E_{t-1}Y_t$. Denoting the forecast $y^e_{t+1} = E_t y_{t+1}$, under rational expectation, equation (19) implies the model of section 2 for y_t (see the appendix for a derivation):

$$y_t = \beta y^e_{t+1} + x_t \tag{20}$$

where the forcing variable is defined as $x_t = \theta \Delta z_t + \beta r_t$, with $\theta = \left((1-\beta)\gamma_1' + \gamma_2'\right)$. x_t can be assumed to be weakly dependent and β is typically close to unity.

In the following two subsections, we provide simulations of the processes involved in the Campbell-Shiller and Engel-West models under adaptive learning. Specifically, we assume x_t to be univariate $i.i.d$ and generate draws thereof from a standard normal distribution. Then, from a random initialization y^e_1, we generate a sample of T observations of $\left(y_t, y^e_{t+1}\right)$ using the algorithms of section 2.2.2 and expression (20).[14] In the following, we show analytically and using the simulated processes how some standard empirical puzzles can be explained under adaptive learning. In the simulations, we use parameter values for the learning algorithm similar to those of section 4. This ensures that the learning window is long relative to the sample size.

5.1 Campbell-Shiller model and predictive regressions

The log-linearized Campbell-Shiller (1988) model can be written as equation (19), where Y_t is log price, z_t log dividends, r_t is excess return and the parameters (γ_1, γ_2) are equal to $(1, 0)$. The log linearization constant β was estimated over 1926-1994 by Campbell, Lo and MacKinlay (1996, chapter 7, p. 261) for asset prices at .96 in annual data and .997 in monthly data.

Campbell and Shiller (1987) define the spread $S_t = Y - z_t$ which they show to be equal to $\beta/(1-\beta)\, y^e_{t+1}$. From this identity, we generate S_t and $\Delta z_t = (y_t - \Delta S_t)$. Finally, we retrieve $r_t = \beta^{-1}\left(x_t - (1-\beta)\Delta z_t\right)$.

Since Fama and Schwert (1977) and Rozeff (1984), it has become standard practice to regress the excess return r_t on lags of log dividend-price ratio $z_t - Y_t$. Such so-called predictive

[14] We set σ_0^2 large enough to be considered almost infinite.

T	$\bar{g}$	γ_0	Mean of $\widehat{\gamma}$ (s.d.)	Rej. Prob. $\mathsf{H}_0:\gamma=0$	Rej. Prob. $\mathsf{H}_0:\gamma=\gamma_0$	Mean of $\widehat{d}$ for S_t	Mean of $\widehat{d}$ for r_t	Mean of $\widehat{d}$ for y_t
Campbell-Shiller								
250	0.00		0.09 (.26)	18.7		0.36 (0.13,0.55)	0.01 (−0.23,0.22)	0.24 (−0.10,0.45)
	0.03	.0017	0.72 (.56)	30.8	30.6	0.50 (0.29,0.69)	−0.00 (−0.23,0.20)	0.09 (−0.14,0.30)
	0.10	.0018	0.27 (.19)	35.0	34.3	0.49 (0.27,0.69)	−0.01 (−0.24,0.20)	0.25 (−0.01,0.46)
400	0.00		0.09 (.25)	20.9		0.37 (0.15,0.53)	0.01 (−0.18,0.18)	0.24 (−0.06,0.43)
	0.03	.0017	0.47 (.36)	33.3	33.0	0.50 (0.32,0.66)	−0.00 (−0.19,0.17)	0.13 (−0.07,0.31)
	0.10	.0018	0.18 (.12)	36.0	35.1	0.49 (0.31,0.66)	−0.00 (−0.19,0.16)	0.29 (0.10,0.47)
Engel-West				$\mathsf{H}_0:\gamma=1$	$\mathsf{H}_0:\gamma=\gamma_0$			
250	0.00		-0.07 (.11)	99.6		0.37 (0.14,0.52)	−0.01 (−0.23,0.21)	0.25 (−0.09,0.46)
	0.03	0.98	0.29 (.58)	32.6	30.5	0.50 (0.29,0.69)	−0.00 (−0.23,0.20)	0.10 (−0.14,0.31)
	0.10	0.98	0.73 (.20)	30.5	34.0	0.50 (0.28,0.69)	−0.00 (−0.24,0.20)	0.27 (0.03,0.48)
400	0.00		-0.05 (.11)	99.8		0.38 (0.15,0.53)	0.01 (−0.19,0.17)	0.26 (−0.06,0.44)
	0.03	0.98	0.54 (.36)	35.0	31.0	0.50 (0.32,0.66)	0.00 (−0.19,0.17)	0.13 (−0.06,0.32)
	0.10	0.98	0.82 (.13)	42.2	30.6	0.50 (0.31,0.66)	−0.00 (−0.19,0.17)	0.31 (0.11,0.48)

Table 3: The table records parameters and simulated statistics for the two examples of estimated predictive regression (Campbell-Shiller, with $\beta = 0.96$) and forward premium anomaly (Engel-West, with $\beta = 0.98$) for sample sizes of 250 and 400 observations. $\widehat{\gamma}$ is the estimator of the parameter of interest in the regression (which takes the value γ_0 under CGLS) and (s.d.) denotes its Monte Carlo standard deviation. Rej. Prob. refers to the Monte Carlo two-sided rejection probability (using asymptotic critical values) of the null that γ is respectively zero (Campbell-Shiller) or unity (Engel-West) together with the null that it takes value γ_0. $\widehat{d}$ is the estimator of the fractional integration order d using a GPH log periodogram regression; the intervals in parentheses underneath report the 2.5% en 97.5% quantiles of the distribution of the estimators.

regressions have often shown that r_t is predictable but the evidence is questionable (see Stambaugh, 1999). Campbell and Yogo (2006) show that when S_t is highly persistent, in the sense that it exhibits a near unit root, the OLS estimator of its coefficient in a predictive regression, say γ, is biased above zero and t-statistics over-reject. In particular, consider the regression model:

$$r_t = c - \gamma S_{t-1} + e_t. \tag{21}$$

In a simulation, Campbell and Yogo show that the t-test empirical size can be as high as 27.2% for a sample of 250 observations. Since S_t is proportional to y^e_{t+1}, then it will be highly persistent under adaptive learning, especially if agents use long window learning.

In order to derive the properties of a predictive regression, let us express r_t as a function of S_{t-1}. Using the model identities, $r_t = \beta^{-1}(x_t - (1-\beta)\Delta z_t)$ can be written as $r_t = y_t - \beta^{-1}(1-\beta)S_{t-1}$. Under learning, the ALM is given by expression (9) where $a_t = y^e_{t+1} = \beta^{-1}(1-\beta)S_{t-1}$, i.e.

$$y_t = \frac{(1-g_t)(1-\beta)}{1-\beta g_t} S_{t-1} + \frac{x_t}{1-\beta g_t}.$$

So, it follows that

$$r_t = -\frac{(1-\beta)^2}{\beta(1-\beta g_t)} S_{t-1} + \frac{x_t}{1-\beta g_t}$$

where x_t is $i.i.d$ by assumption. Hence, the true value of the coefficient on the lag of the log-dividend price ratio $z_{t-1} - Y_{t-1} = -S_{t-1}$ in the predictive regression (21) is given by

$$\gamma_t = \frac{(1-\beta)^2}{\beta(1-\beta g_t)} > 0. \tag{22}$$

This implies that there is predictability in excess returns, arising from the fact that expectations are not rational. However, when β is close to one, and g_t close to zero, γ_t will be close to zero, so predictability is limited. For example, for the calibrated parameter values in Campbell and Shiller given above, $\gamma_t < 0.002$ under RLS or CGLS with small gain, see Table 3. Despite this, we will see that predictability t tests reject the no-predictability null hypothesis with very high probability. We find that this is primarily due to size distortion rather than power, in line with the results in Campbell and Yogo (2006).

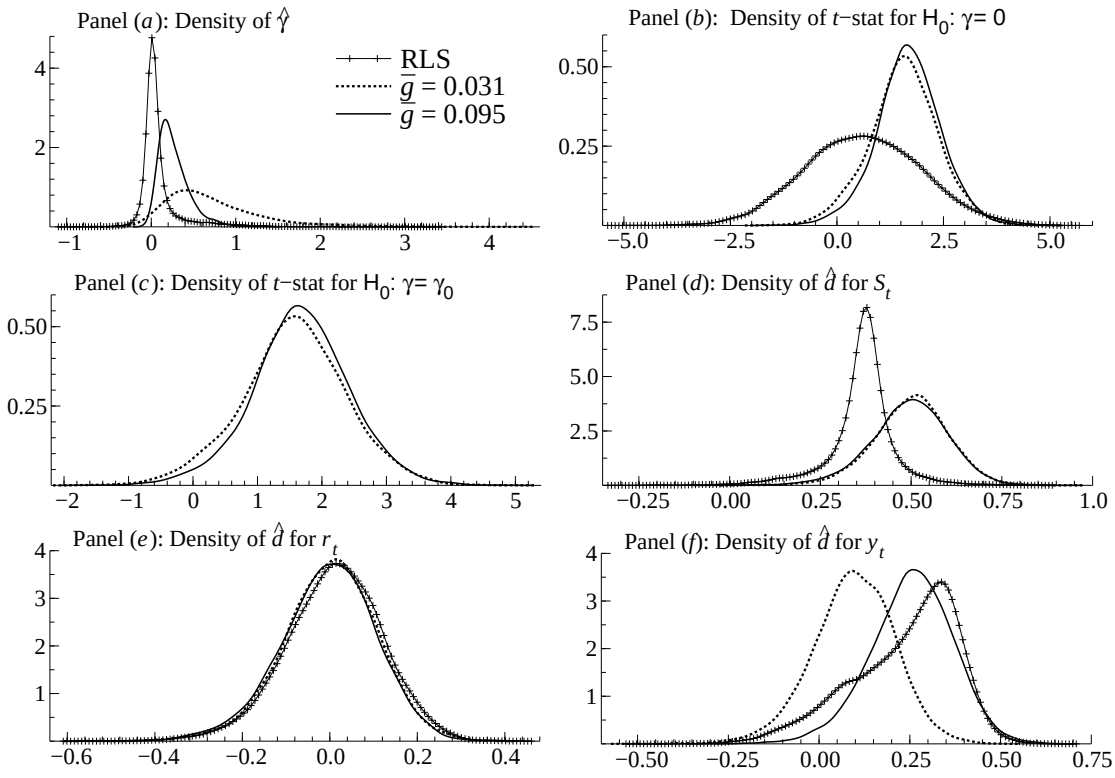

Figure 2: Statistics for the predictive regression derived from the Campbell and Shiller loglinearized asset price model over samples of 250 observations with $\beta = .96$. Panel (a) records the density of regression coefficient estimator; the densities of the corresponding the t-statistic for the nulls of a zero coefficient or the true null (only defined under CGLS) are in panels (b) and (c). Panels (d), (e) and (f) report the densities of the Geweke and Porter-Hudak (1983) estimators of the fractional integration parameter for the spread S_t, the excess return r_t, and the difference y_t in the log price.

Figure 2 presents the sampling distributions of various statistics relating to this model over samples of 250 observations with $\beta = 0.96$ and for different learning algorithms. Panel a reports the OLS estimator $\widehat{\gamma}$ of γ in the predictive regression (21); panels b and c report the t statistics for the hypotheses $\mathsf{H}_0 : \gamma = 0$ and $\mathsf{H}_0 : \gamma = \gamma_0$, where γ_0 is the 'true value' under CGLS, i.e. $\gamma_0 = \frac{(1-\beta)^2}{\beta(1-\beta\bar{g})}$;[15] panels d to f report the densities of the GPH estimators of the long memory parameters for S_t, r_t and y_t, respectively. Table 3 reports summary statistics of those distributions for samples of size 250 and 400.

The t statistic is seen to be centered on strictly positive values with large rejection probabilities of the null $\mathsf{H}_0 : \gamma = 0$ at the 5% level: 19% in the case of RLS learning and more than 30% for CGLS, see table 3. The log-dividend price ratio S_t is persistent, with an estimate for its long memory parameter in the range 0.35-0.5. This is somewhat lower than found by Marinucci and Robinson (2001) (in the range 0.6-0.7 for the non-log ratio), though not significantly so, and it is consistent with estimates based on the annual data in Shiller (1989) updated for 1871-2001 and available from Robert Shiller's website.[16] In particular, the GPH estimate of the degree of long memory for the log dividend-price ratio is 0.28 with 95% confidence interval [-0.01, 0.66]. The corresponding values for the simulated excess returns show much less persistence and no evidence of long memory, which is in line with *inter alia* Lo (1991) and many authors since. Also, the rejection probabilities of the null of no predictive power by Campbell and Yogo (2006) are in line with the simulations results presented in table 3. The change in log prices, y_t, exhibits more persistence than excess returns, and this persistence is increasing in the gain parameter, as our theory predicts. The table and figure also report the t statistics and rejection probabilities when the null hypothesis is true, $\mathsf{H}_0 : \gamma = \gamma_0$, which yields the size of the test. We see that there is significant size distortion, comparable to the results given by Campbell and Yogo (2006), which is symptomatic of the unbalanced nature of the predictive regression. Hence, these results are entirely consistent with the findings of Campbell and Yogo (2006), and they are complementary since they provide an explanation for the persistence in S_t that Campbell and Yogo took as given in their analysis.

[15] RLS learning converges to rational expectations where $E_t y_{t+1} = 0$ so there is no asymptotic predictability.

[16] available at http://www.econ.yale.edu/~shiller/data/chapt26.html

5.2 Exchange rates and the forward premium anomaly

The Forward Premium Anomaly constitutes another puzzling empirical feature that is related to present value models and it has been shown to be understandable in the presence of long range dependence. The puzzle finds its source in the Uncovered Interest Rate Parity (UIP):

$$E_t\left[s_{t+1} - s_t\right] = f_{t,1} - s_t = i_{t,1} - i^*_{t,1}$$

where s_t is the log spot exchange rate, $f_{t,1}$ is the log one-period forward rate, and i_t, i^*_t are the one-period log returns on domestic and foreign risk-free bonds. The UIP under the efficient markets hypothesis has been tested since Fama (1984) as the null $H_0 : (c, \gamma) = (0, 1)$ in the regression:

$$\Delta s_t = c + \gamma\left(f_{t-1,1} - s_{t-1}\right) + \epsilon_t. \tag{23}$$

The anomaly lies in the rejection of H_0 with an estimate $\widehat{\gamma} << 1$, often negative.

Among the possible explanations for the puzzle, data persistence of exogenous origin has been raised inter alia by Engel and West (2005). Baillie and Bollerslev (2000) and Maynard and Phillips (2001) suggest an explanation through fractional integration. There is ample evidence in the literature that changes in log exchange rates and the forward premium exhibit persistence that is less strong than implied by the presence of a pure unit root. These variables have been modeled for instance as near-I(1) or fractionally integrated, see inter alia Cheung (1993), Baillie (1996), Baillie and Bollerslev (1994a, 1994b, 2000), and Engel and West (2005). Maynard and Phillips (2001, tables I and II) in particular showed that the estimated fractional integration orders of the forward premia vary mostly in the range $(.2, .6)$ which is comparable to the values in Baillie and Bollerslev (1994a), between .4 and .7. Regarding the differences in spot log exchange rates, Cheung (1993) estimated integration orders lying within (0,.5) using the GPH estimator and much lower values using Maximum Likelihood techniques, as did Maynard and Phillips. Many explanations have been proposed for this that rely on the persistence of the processes that are modeled either as fractionally integrated (Maynard and Phillips, 2001) or as exhibiting persistent volatility (Baillie and Bollerslev, 2000).

Engel and West (2005) discuss some exchange rate models that can be written in the form (19), with $Y_t = s_t$ and where z_t is the log fundamental. In particular, they discuss a money

income model and a Taylor rule model where the foreign country has an explicit exchange rate target. In the money income model, the discount rate β is function of the interest semi-elasticity of money demand. Using past empirical studies, Engel and West evaluate that it should lie within the range 0.97-0.98. In the Taylor rule model, β relates negatively to the degree of the intervention of foreign monetary authorities to target the exchange rate. Empirical evidence allows the authors to evaluate its range between 0.975 and 0.988.

Under the UIP, the forward premium is given by $f_{t,1} - s_t = y^e_{t+1}$, where $y_t = \Delta s_t$. Using the mean-plus-noise learning model of section 2.2.2 and equation (20), we obtain

$$\Delta s_t = \frac{\beta\,(1-g_t)}{1-\beta g_t}\,(f_{t-1,1} - s_{t-1}) + \frac{x_t}{1-\beta g_t}.$$

Hence, the true value of γ in the regression (23) is

$$\gamma_t = \frac{\beta\,(1-g_t)}{1-\beta g_t} < 1. \tag{24}$$

We see that the hypothesis $\gamma = 1$ clearly does not hold (because expectations are not rational), but with β close to one and g_t close to zero, γ_t can be very close to unity. The persistence of Δs_t and of the forward premium $S_t = f_{t,1} - s_t$ under learning implies that the bias of the OLS estimator $\widehat{\gamma}$ of γ is large and negative in finite samples.

Figure 3 presents the sampling distributions of various statistics relating to this model over samples of 250 observations with $\beta = 0.98$ and for different learning algorithms. Panel a reports the OLS estimator $\hat{\gamma}$ of γ in the regression (23); panels b and c report the t statistics for the hypotheses $\mathsf{H}_0 : \gamma = 1$ and $\mathsf{H}_0 : \gamma = \gamma_0$, where γ_0 is the 'true value' under CGLS, i.e., $\gamma_0 = \frac{\beta(1-\bar{g})}{1-\beta\bar{g}}$;[17] panels d to f report the densities of the GPH estimators of the long memory parameters for $S_t, r_t = s_t - s^e_t$ and y_t, respectively. Table 3 reports summary statistics of those distributions for samples of size 250 and 400.

We observe that the estimated coefficients are indeed lower than unity and that the null that $\gamma = 1$ in equation (23) is rejected with a high probability (over 99% under RLS, 25% with CGLS, see table 3). The forward premium is, according to the simulations, estimated fractionally integrated of order within the range 0.35-0.50 and the change in log spot rates exhibits less persistence (mostly non-significant). Table 3 shows that under RLS learning, $\hat{\gamma}$ is strongly downwardly biased. As with the previous example, the rejection probabilities

[17] Under RLS learning, the UIP holds asymptotically since learning converges to rational expectations.

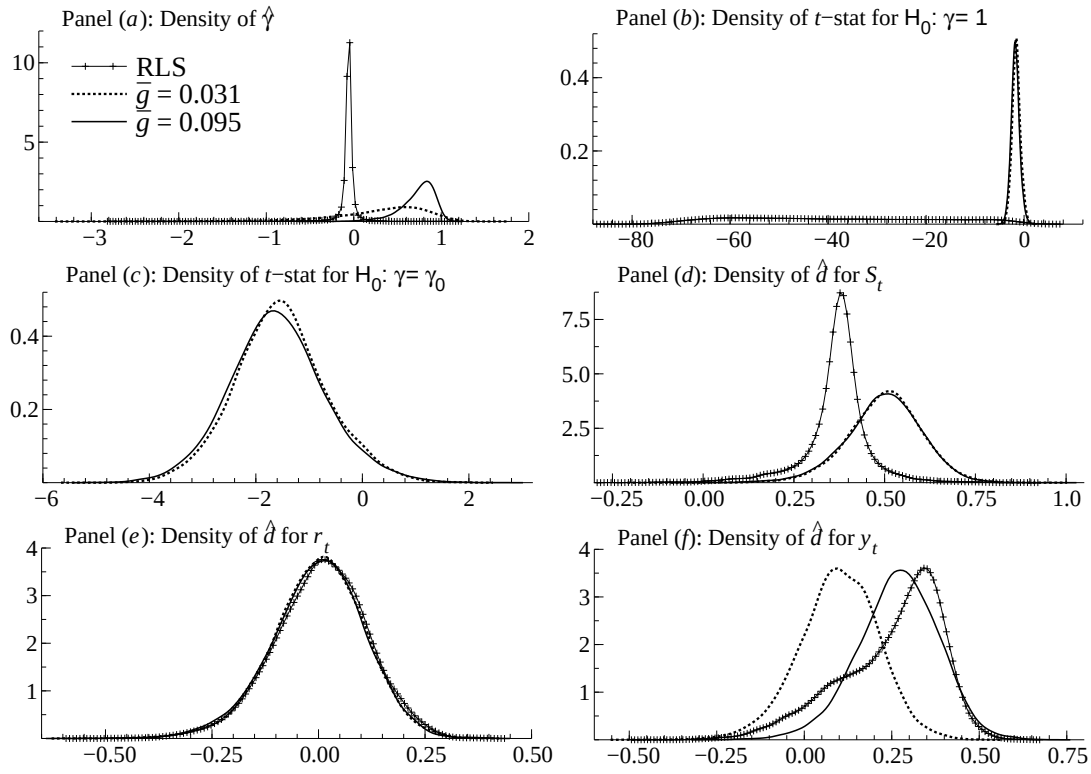

Figure 3: Statistics for the UIP regression derived from the Campbell and Shiller loglinearized asset price model over samples of 250 observations with $\beta = .98$. Panel (a) records the density of regression coefficient estimator; the densities of the corresponding t-statistics for the nulls of a unit coefficient and for the true null (defined only under CGLS) are in panels (b) and (c) respectively. Panels (d), (e) and (f) report the densities of the Geweke and Porter-Hudak (1983) estimators of the fractional integration parameter for the spread S_t, the innovations r_t, and the difference y_t in the log exchange rate.

for the null that $\gamma = 1$ or γ_0 are close (in particular when the gain is low). Hence the high rejection probabilities are not due to high power, but more due to size distortion. Finally, the degree of persistence of change in log spot rates, y_t, although higher than that of the innovation, r_t, is not significant when the gain is low and is in line with the values reported in Cheung (1993) and Maynard and Phillips (2001).

6 Conclusion

We studied the implications of learning in models where endogenous variables depend on agents' expectations. In a prototypical representative-agent forward-looking model, with linear learning algorithms, we found that learning can generate strong persistence. The degree of persistence induced by learning depends on the weight agents place on past observations when they update their beliefs, and on the magnitude of the feedback from expectations to the endogenous variable. In the special case of a long-window learning algorithm known as recursive least squares, long memory arises when the coefficient on expectations is greater than a half. In algorithms with shorter window, long memory provides an approximation to the low-frequency variation of the endogenous variable that can be made arbitrarily accurate in finite samples. Importantly, long memory arises endogenously here, due to the self-referrential nature of the model, without the need for any persistence in the exogenous shocks. This is distinctly different from the behavior of the model under rational expectations, where the memory of the endogenous variable is determined exogenously and the feedback on expectations has no impact. Moreover, our results are obtained without any of the features that have been previously shown in the literature to be associated with long memory, such as structural change, heterogeneity and nonlinearities. Finally, we showed that this property of learning can be used to shed light on some well-known empirical puzzles in present value models.

Appendix

A Proof of proposition 1

We look for a solution $y_t = \sum_{j=0}^{\infty} \psi_j \eta_{t-j}$ that satisfies $y_t = \beta E_t y_{t+1} + x_t$ with $\beta \leq 1$. This implies

$$\sum_{j=0}^{\infty} (\psi_j - \beta\psi_{j+1}) \eta_{t-j} = \sum_{j=0}^{\infty} \theta_j \eta_{t-j}.$$

Identifying the coefficients, it follows that $\psi_j - \beta\psi_{j+1} = \theta_j$ for all $j \geq 0$, so

$$\psi_j = \theta_j + \beta\psi_{j+1} = \sum_{k=j}^{\infty} \beta^{k-j}\theta_k.$$

Hence as $j \to \infty$, $\psi_j \to 0$ and the rate of decay of the (ψ_j) coefficients will be slowest when $\beta = 1$. When $\beta < 1$,

$$\psi_j = O\left(\theta_j \sum_{k=0}^{\infty} \beta^k\right) = O(\theta_j),$$

so $\left|\sum_{j=0}^{\infty} \psi_j\right| < \infty$ if $\left|\sum_{j=0}^{\infty} \theta_j\right| < \infty$. We then use theorem 3.11 of Phillips and Solo (1992) who show that y_t then satisfies a Central Limit Theorem.

If $\beta = 1$, then $\psi_j = \sum_{k=j}^{\infty} \theta_k$ so

$$\sum_{j=0}^{\infty} \psi_j = \sum_{j=0}^{\infty} (j+1)\theta_j$$

and the result follows assuming that $\sum_{j=0}^{\infty} j|\theta_j| < \infty$.

B Preliminary lemmas

We introduce the following two lemmas which we will use in several instances in the proofs.

Lemma 7 *Let $\kappa(L) = \sum_{j=0}^{\infty} \kappa_j L^j$ with $\kappa_j \sim c_\kappa j^{\delta_\kappa - 2}$ as $j \to \infty$, for $c_\kappa > 0$ and $\delta_\kappa \in (0,1)$. Then, there exist strictly positive scalars $(c_\kappa^*, c_\kappa^{**})$ such that*

$$\begin{aligned}
&\mathrm{Re}\left(\kappa\left(e^{i\omega}\right) - 1\right) \underset{\omega \to 0^+}{=} -c_\kappa^* \omega^{1-\delta_\kappa} + o\left(\omega^{1-\delta_\kappa}\right), \\
&\left|\kappa\left(e^{i\omega}\right) - 1\right|^2 \underset{\omega \to 0^+}{=} c_\kappa^{**} \omega^{2(1-\delta_\kappa)} + o\left(\omega^{2(1-\delta_\kappa)}\right).
\end{aligned}$$

Proof. For $\delta_\kappa \in (0,1)$, $\delta_\kappa - 2 \in (-2,-1)$. Yong (1974), theorems III-24 and -27, show that for $\delta_\kappa \in (0,1)$, if there exists a function S slowly varying at infinity such that $a_j = j^{\delta_\kappa - 2} S(j)$, then

$$\sum_{j=1}^{\infty} a_j \cos(j\omega) - \sum_{j=1}^{\infty} a_j \underset{\omega \to 0^+}{\sim} \frac{\pi}{2\Gamma(2-\delta_k)\cos\frac{(2-\delta_k)\pi}{2}} \omega^{1-\delta_\kappa} S\left(\frac{1}{\omega}\right)$$
$$\sum_{j=1}^{\infty} a_j \sin(j\omega) \underset{\omega \to 0^+}{\sim} \frac{\pi}{2\Gamma(2-\delta_k)\sin\frac{(2-\delta_k)\pi}{2}} \omega^{1-\delta_\kappa} S\left(\frac{1}{\omega}\right).$$

Define $S(x) = \kappa_{\lceil x \rceil} / \lceil x \rceil^{\delta_\kappa - 2}$, where $\lceil x \rceil$ is the integer part of x_t. Then as $x \to \infty$ and for $\lambda > 0$,

$$S(\lambda x)/S(x) = \frac{\kappa_{\lceil \lambda x \rceil}}{\kappa_{\lceil x \rceil}} \frac{\lceil x \rceil^{\delta_\kappa - 2}}{\lceil \lambda x \rceil^{\delta_\kappa - 2}} \to 1,$$

so S is slowly varying with $S\left(\frac{1}{\omega}\right) \to c_\kappa$ as $\omega \to 0$. This implies, using $\kappa_j = j^{\delta_\kappa - 2} S(j)$ and $\sum_{j=1}^{\infty} \kappa_j = 1$, that

$$\kappa\left(e^{-i\omega}\right) - 1 \underset{\omega \to 0^+}{\sim} \frac{\pi c_\kappa}{2\Gamma(2-\delta_\kappa)} \left[-\frac{1}{\cos\frac{\pi\delta_\kappa}{2}} + i \frac{1}{\sin\frac{\pi\delta_\kappa}{2}} \right] \omega^{1-\delta_\kappa},$$

i.e. the result holds for $\mathrm{Re}\left(\kappa\left(e^{i\omega}\right) - 1\right)$ setting $c_\kappa^* = \frac{\pi c_\kappa}{2\Gamma(2-\delta_\kappa)} \frac{1}{\cos\frac{\pi\delta_\kappa}{2}} > 0$. Also, using

$$\left| \frac{-1}{\cos\frac{\pi\delta_\kappa}{2}} + i\frac{1}{\sin\frac{\pi\delta_\kappa}{2}} \right|^2 = \left(\cos\frac{\pi\delta_\kappa}{2} \sin\frac{\pi\delta_\kappa}{2} \right)^{-2} = \left(\frac{\sin\pi\delta_\kappa}{2} \right)^{-2},$$

and $\Gamma(1+z) = z\Gamma(z)$ together with $\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin\pi z}$, we obtain

$$\left|1 - \kappa\left(e^{i\omega}\right)\right|^2 \sim \frac{c_\kappa^2 \Gamma(\delta_\kappa)^2}{(1-\delta_\kappa)^2} \omega^{2(1-\delta_\kappa)}, \tag{25}$$

and $c_\kappa^{**} = \frac{c_\kappa^2 \Gamma(\delta_\kappa)^2}{(1-\delta_\kappa)^2} > 0$. ■

Lemma 8 *Consider the model $y_t = \beta y_{t+1}^e + x_t$, with $y_{t+1}^e = \kappa(L) y_t$. Suppose x_t satisfies assumption B, and that the constant learning algorithm $\kappa(\cdot)$ satisfies assumption A with $\delta_\kappa \in (0,1)$. Then the spectral density of y_t is finite at the origin $f_y(0) < \infty$ and admits an upward vertical asymptote: there exists $c_f > 0$ such that*

$$f_y'(0) \underset{\omega \to 0}{\sim} -c_f \omega^{-\delta_\kappa}. \tag{26}$$

Proof. Consider

$$f_y(\omega) - f_y(0) = \frac{(f_x(\omega) - f_x(0))(1-\beta)^2}{(1-\beta)^2 |1-\beta+\beta(1-\kappa(e^{-i\omega}))|^2} - f_x(0) \frac{2\beta(1-\beta)\operatorname{Re}\left[1-\kappa\left(e^{-i\omega}\right)\right] + \beta^2 \left|1-\kappa\left(e^{-i\omega}\right)\right|^2}{(1-\beta)^2 |1-\beta+\beta(1-\kappa(e^{-i\omega}))|^2},$$

since

$$\left|1-\beta+\beta\left(1-\kappa\left(e^{-i\omega}\right)\right)\right|^2 = (1-\beta)^2 - 2\beta(1-\beta)\operatorname{Re}\left[\kappa\left(e^{-i\omega}\right)-1\right] - \beta^2\left|\kappa\left(e^{-i\omega}\right)-1\right|^2.$$

Now if $\delta_\kappa > 0$, under constant learning, $\kappa_j \sim c_\kappa j^{\delta_\kappa - 2}$ for some $c_\kappa > 0$. Lemma 7 implies that there exist $c_\kappa^*, c_\kappa^{**}$ such that

$$f_y(\omega) - f_y(0) \underset{\omega\to 0^+}{\sim} \frac{\frac{(1-\beta)^2}{2} f_x''(\omega)\omega^2 - 2\beta(1-\beta) f_x(0) c_\kappa^* \omega^{1-\delta_\kappa} + \beta^2 f_x(0) c_\kappa^{**} \omega^{2(1-\delta_\kappa)}}{(1-\beta)^4} \tag{27}$$

$$\underset{\omega\to 0^+}{\sim} -\frac{2\beta f_x(0) c_\kappa^*}{(1-\beta)^3}\omega^{1-\delta_\kappa}.$$

We first note that by definition of the population spectrum

$$f_y(\omega) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma_k e^{-i\omega k} = \frac{1}{2\pi}\left\{\gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \cos\omega k\right\},$$

where $\gamma_k = Cov(y_t, y_{t-k})$ is symmetric since y_t is stationary. We assume for now that γ_k is of bounded variation, then, for $\omega \neq 0$, the series $\sum_{k=1}^n \gamma_k \cos\omega k$ converges uniformly as $n \to \infty$ (see Zygmund, 1935, section 1.23). It follows that the derivative of f_y satisfies:

$$f_y'(\omega) = -\frac{1}{\pi}\sum_{k=1}^{\infty} k\gamma_k \sin k\omega. \tag{28}$$

We now use theorem III-11 of Yong (1974) who works under the assumption that $\{a_k\}_{k\in\mathbb{N}}$ is a sequence of positive numbers that is quasi-monotonically convergent to zero (i.e. $a_k \to 0$ and there exist $M \geq 0$, such that $a_{k+1} \leq a_k\left(1+\frac{M}{k}\right)$ for all $k \geq k_0(M)$) and that $\{a_k\}$ is also of bounded variation, i.e. $\sum_{k=1}^{\infty}|\Delta a_k| < \infty$. The theorem states that for $a \in (0,1)$, $a_k \sim k^{-a}S^*(k)$ as $k \to \infty$,with S^* slowly varying, if and only if

$$\sum_{k=1}^{\infty} a_k \sin k\omega \sim \frac{\pi}{2\Gamma(a)\sin\frac{\pi a}{2}}\omega^{a-1}S^*\left(\frac{1}{\omega}\right) \text{ as } \omega \to 0^+.$$

We apply this theorem to (28), using expression (27) that $f'_y(\omega) \underset{\omega\to 0^+}{\sim} -\frac{2\beta f_x(0) c^*_\kappa}{(1-\beta)^3}\omega^{-\delta_\kappa}$

$$-\frac{1}{\pi}\sum_{k=1}^{\infty} k\gamma_k \sin k\omega \underset{\omega\to 0^+}{\sim} -\frac{2\beta f_x(0) c^*_\kappa}{(1-\beta)^3}\omega^{-\delta_\kappa}.$$

We let $a = 1-\delta_\kappa$ in the theorem of Yong above, defining

$$a_k = \frac{k\gamma_k}{\pi}\frac{(1-\beta)^3}{2\beta f_x(0) c^*_\kappa}.$$

This implies that $a_k \sim \frac{\pi}{2\Gamma(a)\sin\frac{\pi a}{2}} k^{-(1-\delta_\kappa)}$, with $\Gamma(1-\delta_\kappa)\sin\frac{\pi(1-\delta_\kappa)}{2} = \frac{\pi}{2\Gamma(\delta_\kappa)\sin\frac{\pi\delta_\kappa}{2}}$, i.e.

$$\gamma_k \sim \frac{2\pi\beta f_x(0) c^*_\kappa \Gamma(\delta_\kappa)\sin\frac{\pi\delta_\kappa}{2}}{(1-\beta)^3} k^{-(2-\delta_\kappa)}. \tag{29}$$

To apply theorem III-11 of Yong (1974), we need to check that a_k thus defined is a quasi-monotonic sequence with bounded variation. The first holds since $a_{k+1}/a_k \sim (1+1/k)^{-(1-\delta_\kappa)} < 1$, so choose M such that $a_{k+1}/a_k < 1$ for $k > M$. Also $k\gamma_k$ is clearly of bounded variation since it is asymptotically positive and

$$\Delta(k\gamma_k) = O\left(k^{-(2-\delta_\kappa)}\right)$$

is summable. Finally, we need to check the uniform convergence condition in Zygmund (1935): $|\Delta\gamma_k| = O\left(k^{-(3-\delta_\kappa)}\right)$ so γ_k is of bounded variation. ■

C Proof of theorem 2

Substitute (4) into (1) to get

$$y_t = \beta\sum_{j=0}^{t-1}\kappa_j y_{t-j} + \beta\zeta_t + x_t,$$

and define $\kappa^*(L) = 1-\kappa(L) = \sum_{j=0}^{\infty}\kappa^*_j L^j$ so

$$(1-\beta) y_t + \beta\sum_{j=0}^{t-1}\kappa^*_j y_{t-j} = x_t + \beta\zeta_t.$$

Summing yields

$$\sum_{t=1}^{T}\left((1-\beta) - \beta\sum_{j=0}^{t-1}\kappa^*_j\right) y_{T-t+1} = \sum_{t=1}^{T}(x_t + \beta\zeta_t). \tag{30}$$

The left-hand side of the previous equation shows that the magnitude of $\sum_{t=1}^{T} y_t$ depends on the limit of $(1-\beta)/\sum_{j=0}^{T-1}\kappa_j^*$. Since $\kappa^*(1)=0$, $\kappa_j \sim c_\kappa j^{\lambda-2}$ with $\lambda<1$ implies $\kappa_j^* \sim -c_\kappa j^{\lambda-2}$ and $\sum_{j=0}^{T-1}\kappa_j^* \sim \frac{c_\kappa}{1-\lambda}T^{\lambda-1}$. Under assumption A, the previous expressions hold with $\lambda=\delta_\kappa$ when $\delta_\kappa \in (0,1)$ and for $\lambda<0$ if $\delta_\kappa=0$.

We consider the three cases in turn. Defining $y_t^- = y_t 1_{\{t\leq 0\}}$, we made the following assumptions about ζ_t :

$$\text{if } \nu \leq 1-\delta_\kappa : \sum_{t=1}^{T}\zeta_t = o_p\left(\sum_{t=1}^{T} x_t\right); \tag{31}$$

$$\text{if } \nu > 1-\delta_\kappa : \begin{cases} \zeta_t = \kappa(L)\, y_t^-, & \text{if } \delta_\kappa \in \left(\frac{1}{2},1\right); \\ \Delta\zeta_t = (1-L)\,\kappa(L)\, y_t^-, & \text{if } \delta_\kappa \in \left(0,\frac{1}{2}\right). \end{cases} \tag{32}$$

First if $(1-\beta)/\sum_{j=0}^{T-1}\kappa_j^* \to \infty$, i.e. $\nu<1-\delta_\kappa$, or $\delta_\kappa=0$, then

$$\sum_{t=1}^{T}\left((1-\beta)-\beta\sum_{j=0}^{t-1}\kappa_j^*\right) y_{T-t+1} = (1-\beta)\sum_{t=1}^{T} y_t + o_p\left((1-\beta)\sum_{t=1}^{T} y_t\right),$$

hence, with assumption (31),

$$\sum_{t=1}^{T} y_t = O_p\left((1-\beta)^{-1}\sum_{t=1}^{T} x_t\right) = O_p\left(T^{\frac{1}{2}+\nu}\right)$$

$$var\left(T^{-1/2}\sum_{t=1}^{T} y_t\right)^{1/2} = O\left(T^{\nu}\right).$$

Consider now the case where $(1-\beta)/\sum_{j=0}^{T-1}\kappa_j^* \to 0$ i.e. $\nu>1-\delta_\kappa$. First assume that $\delta_\kappa \in \left(\frac{1}{2},1\right)$. Define $z_t = [\kappa^*(L)]^{-1} x_t$ with spectral density

$$f_z(\omega) = \frac{f_x(\omega)}{\left|1-\kappa\left(e^{-i\omega}\right)\right|^2}.$$

Using lemma 7, as $\omega \to 0$

$$f_z(\omega) \sim \frac{f_x(0)}{c_\kappa^{**}}\omega^{-2(1-\delta_\kappa)}. \tag{33}$$

Beran (1994, theorem 2.2 p. 45) shows that (33) implies that

$$var\left(\sum_{t=1}^{T} z_t\right) = O\left(T^{1+2(1-\delta_\kappa)}\right).$$

Notice that the proof is in the appendix of Beran (1989) and relies on showing that $f_z(\omega)$ can be written as $\left|1-e^{-i\omega}\right|^{-2(1-\delta_\kappa)} S(1/\omega)$ where S is slowly varying at infinity.

Under assumption (32) and noting that $\kappa(L) y_t^- = (\kappa(L)-1) y_t^-$, expression (30) rewrites

$$\sum_{t=1}^{T}\left((1-\beta)-\beta\sum_{j=0}^{t-1}\kappa_j^*\right) y_{T-t+1} - \beta\sum_{t=0}^{\infty}\sum_{j=t+1}^{t+T}\kappa_j y_{-t} = \sum_{t=1}^{T} x_t.$$

Since $(1-\beta) = o\left(\sum_{j=0}^{T-1}\kappa_j^*\right)$, it follows that, denoting $y_t^+ = y_t - y_t^-$,

$$\begin{aligned}
&\sum_{t=1}^{T}\left((1-\beta)-\beta\sum_{j=0}^{t-1}\kappa_j^*\right) y_{T-t+1} - \beta\sum_{t=0}^{\infty}\sum_{j=t+1}^{t+T}\kappa_j y_{-t} \\
&= -\beta\left[\sum_{t=1}^{T}\left(\sum_{j=0}^{t-1}\kappa_j^*\right) y_{T-t+1} + \sum_{t=0}^{\infty}\sum_{j=t+1}^{t+T}\kappa_j y_{-t}\right] + o_p\left(\sum_{t=1}^{T}\sum_{j=0}^{t-1}\kappa_j^* y_{T-t+1}\right) \\
&= \sum_{t=1}^{T}(1-\kappa(L)) y_t + o_p\left(\sum_{t=1}^{T}(1-\kappa(L)) y_t^+\right).
\end{aligned}$$

Hence, using $\sum_{t=1}^{T} x_t = \sum_{t=1}^{T}(1-\kappa(L)) z_t$,

$$\begin{aligned}
\sum_{t=1}^{T}(1-\kappa(L)) y_t + o_p\left(\sum_{t=1}^{T}(1-\kappa(L)) y_t^+\right) &= \sum_{t=1}^{T} x_t \\
\sum_{t=1}^{T}(1-\kappa(L))(y_t - z_t) + o_p\left(\sum_{t=1}^{T}(1-\kappa(L)) y_t^+\right) &= 0 \\
\sum_{t=1}^{T}(y_t - z_t) + o_p\left(\sum_{t=1}^{T} y_t\right) &= 0
\end{aligned}$$

i.e.

$$\sqrt{var\left(T^{-1/2}\sum_{t=1}^{T} y_t\right)} = O\left(T^{1-\delta_\kappa}\right). \tag{34}$$

Finally, if $\delta_\kappa \in (0, 1/2)$, defining $\Delta z_t = [\kappa^*(L)]^{-1}\Delta x_t$, and following the previous steps under assumption (32) starting from $(1-\beta\kappa(L))\Delta y_t = \Delta x_t + \beta\Delta\zeta_t$ leads to

$$\sum_{t=1}^{T}\Delta(y_t - z_t) + o_p\left(\sum_{t=1}^{T}\Delta y_t\right) = 0.$$

The result by Beran (1989) regarding the magnitude of $var\left(\sum_{t=1}^{T}\Delta z_t\right)$ cannot be used here for $(1-\delta_\kappa) \in \left(\frac{1}{2}, 1\right)$. Yet, the spectral density of Δz_t satisfies

$$f_{\Delta z}(\omega) \sim \frac{f_x(0)}{c_\kappa^{**}}\omega^{2\delta_\kappa},$$

which implies (see Lieberman and Phillips, 2008) that $\gamma_{\Delta z}(k) \sim k^{-2\delta_\kappa - 1}$. Also $f_{\Delta z}(0) = 0$ so $\gamma_{\Delta z}(0) + 2\sum_{k=1}^{\infty} \gamma_{\Delta z}(k) = 0$. The long run variance of Δz_t is hence such that

$$\begin{aligned} var\left(T^{-1}\sum_{t=1}^{T}\Delta z_t\right) &= \gamma_{\Delta z}(0) + 2T^{-1}\sum_{k=1}^{T-1}(T-k)\gamma_{\Delta z}(k) \\ &= \left(\gamma_{\Delta z}(0) + 2\sum_{k=1}^{T-1}\gamma_{\Delta z}(k)\right) - 2T^{-1}\sum_{k=1}^{T-1}k\gamma_{\Delta z}(k) \\ &= -\sum_{k=T}^{\infty}\gamma_{\Delta z}(k) - 2T^{-1}\sum_{k=1}^{T-1}k\gamma_{\Delta z}(k) \\ &= O\left(T^{-2\delta_\kappa}\right) + O\left(T^{-1+1-2\delta_\kappa}\right) \\ &= O\left(T^{-2\delta_\kappa}\right). \end{aligned} \tag{35}$$

The last remaining case we must consider is when $\nu = 1 - \delta_\kappa$, starting with assuming $\delta_\kappa \neq 0$. Then $(1-\beta)$ and $\sum_{j=0}^{t-1}\kappa_j^*$ are of comparable magnitude and we cannot neglect either. We therefore use a direct proof: Brillinger (1975, theorem 5.2.1) shows that if the covariances of y_t are summable and $E(y_t) = 0$,

$$\frac{var\left(T^{-1}\sum_{t=1}^{T}y_t\right)}{f_y(0)} = (2\pi T)^{-1}\int_{-\pi}^{\pi}\frac{\sin^2(T\omega/2)}{\sin^2(\omega/2)}\frac{f_y(\omega)}{f_y(0)}d\omega, \tag{36}$$

where $f_y(\omega)$ is the spectral density of y_t. The function $\left[\frac{\sin(T\omega/2)}{\sin(\omega/2)}\right]^2$ achieves its maximum over $[-\pi,\pi]$ at zero where its value is T^2. As $T \to \infty$ it remains bounded for all $\omega \neq 0$. It is therefore decreasing in ω in a neighborhood of 0^+. Lemma 8 shows that $f_y(\omega)$ is also decreasing in such a neighborhood and $\frac{f_y(\omega)}{f_y(0)}$ is bounded at all T. Both functions being positive, their product is also decreasing in ω in a neighborhood of 0^+; it is in addition continuous, even and differentiable at all $\omega \neq 0$. As $T \to \infty$, the integrand of (36) presents a pole at the origin and its behavior in the neighborhood of zero governs the magnitude of the integral. Since the integrand achieves its local maximum at zero, we can restrict our analysis to a neighborhood thereof, $[0, \theta_T]$ with $\theta_T = o\left(T^{-1}\right)$ since $\frac{\sin^2(T\theta_T/2)}{\sin^2(\theta_T/2)}\frac{f_y(\omega)}{f_y(0)}$ remains bounded as $T \to \infty$ for any sequence θ_T such that $T\theta_T \not\to 0$.

Let $\varepsilon > 0$ and $\beta = 1 - c_\beta T^{-\nu}$, we develop the integrand of (36) about the origin, provided $T^\nu \theta_T^{1-\delta_\kappa} = \left(T^{\nu/(1-\delta_\kappa)}\theta_T\right)^{1-\delta_\kappa} = o(1)$, i.e. if $\nu \leq 1 - \delta_\kappa$. This yields for the integral over

$[0, \theta_T]$:

$$
\begin{aligned}
&(2\pi T)^{-1} \int_0^{\theta_T} \left(T^2 \left(1 - \frac{1}{3} \left(T^2 - 1 \right) \omega^2 + o\left(T^2 \omega^2 \right) \right) \right) \left(1 - c_V T^{\nu} \omega^{1-\delta_\kappa} + o\left(T^{\nu} \omega^{1-\delta_\kappa} \right) \right) d\omega \\
&= \frac{T}{2\pi} \left[\theta_T - \frac{1}{9} \left(T^2 - 1 \right) \theta_T^3 - \frac{c}{2 - \delta_\kappa} T^{\nu} \theta_T^{2-\delta_\kappa} + \frac{c_V}{3\left(4 - \delta_\kappa \right)} \left(T^2 - 1 \right) T^{\nu} \theta_T^{4-\delta_\kappa} \right] \\
&= \frac{T}{2\pi} \left[T^{-(1+\varepsilon)} - \frac{T^2 - 1}{9} T^{-3(1+\varepsilon)} - \frac{c_V}{2 - \delta_\kappa} T^{\nu - (2-\delta_\kappa)(1+\varepsilon)} + \frac{c_V \left(T^2 - 1 \right)}{3\left(4 - \delta_\kappa \right)} T^{\nu} T^{-(4-\delta_\kappa)(1+\varepsilon)} \right] \\
&\sim \frac{1}{2\pi} \left[T^{-\varepsilon} - \frac{1}{9} T^{-3\varepsilon} - \frac{c_V}{2 - \delta_\kappa} T^{\nu - (1-\delta_\kappa) - (2-\delta_\kappa)\varepsilon} + \frac{c_V}{3\left(4 - \delta_\kappa \right)} T^{\nu - (1-\delta_\kappa) - (4-\delta_\kappa)\varepsilon} \right], \quad (37)
\end{aligned}
$$

where c_V is implicitly defined from lemma 8. Expression (37) shows that if $\nu \leq 1 - \delta_\kappa$ the integral over $[0, \theta_T]$ – and hence that over $[-\pi, \pi]$ – remains bounded in the neighborhood of the origin and hence $\frac{var\left(T^{-1} \sum_{t=1}^{T} y_t \right)}{f_y(0)} = O\left(1 \right)$, with $f_y\left(0 \right) = \left(1 - \beta \right)^{-2} f_x\left(0 \right)$, i.e.

$$
var\left(T^{-1} \sum_{t=1}^{T} y_t \right) = O\left(T^{2\nu} \right). \quad (38)
$$

Now when $\left(\delta_\kappa, \nu \right) = \left(0, 1 \right)$, assumption A.3 implies that $0 < \kappa'\left(1 \right) = \sum_{j=1}^{\infty} j \kappa_j < \infty$. By lemma 2.1 of Phillips and Solo (1992), there exists a polynomial $\widetilde{\kappa}$ such that

$$
\kappa\left(L \right) = 1 - \left(1 - L \right) \widetilde{\kappa}\left(L \right),
$$

with $\widetilde{\kappa}\left(1 \right) < \infty$. $\widetilde{\kappa}\left(L \right) = \left(1 - L \right)^{-1} \left(1 - \kappa\left(L \right) \right)$ so the roots of $\widetilde{\kappa}$ coincide with the values z such that $\kappa\left(z \right) = 1$, except at $z = 1$ for which $\widetilde{\kappa}\left(1 \right) = \kappa'\left(1 \right) > 0$ (by L'Hospital's rule). $\kappa\left(z \right) = 1$ implies that the roots of $\widetilde{\kappa}\left(L \right)$ lie outside the unit circle ($\kappa\left(z \right) < \kappa\left(1 \right) = 1$ for $|z| \leq 1$, $z \neq 1$) and the process $\widetilde{x}_t$ defined by $\widetilde{\kappa}\left(L \right) \widetilde{x}_t = x_t$ is $\mathsf{I}(0)$ with differentiable spectral density at the origin. Hence y_t satisfies the near-unit root definition of Phillips (1987):

$$
\left(1 - \beta L \right) y_t = \widetilde{x}_t,
$$

and the result follows.

D Proof of theorem 3

We present in turn the proofs for the spectral density and the autocorrelation

D.1 Spectral density

We consider the behavior of the spectral density of y_t about the origin under the assumption that $\kappa_j \sim c_\kappa j^{\delta_\kappa - 2}$ so define $(c_\kappa^*, c_\kappa^{**})$ as in lemma 7. Let $\beta = 1 - c_\beta T^{-\nu}$, $\nu \in [0,1]$. As $\omega \to 0^+$, the spectral density of f_y is, for $\delta_\kappa \in (1/2, 1)$:

$$f_y(\omega) = \frac{f_x(\omega)}{|1 - \beta\kappa(e^{-i\omega})|^2} = \frac{f_x(\omega)}{|1 - \beta + \beta(1 - \kappa(e^{-i\omega}))|^2}, \tag{39}$$

which implies

$$\begin{aligned} & f_y(\omega) \\ &= \frac{f_x(\omega)}{(1-\beta)^2 - 2\beta c_\kappa^*(1-\beta)\omega^{1-\delta_\kappa} + \beta^2 c_\kappa^{**}\omega^{2(1-\delta_\kappa)} + o((1-\beta)\omega^{1-\delta_\kappa}) + o\left(\omega^{2(1-\delta_\kappa)}\right)}. \end{aligned} \tag{40}$$

Hence when $\delta_\kappa \in (0, 1/2)$:

$$f_{\Delta y}(\omega) = \frac{f_x(\omega)\left(\omega^2 + o\left(\omega^2\right)\right)}{(1-\beta)^2 - 2\beta c_\kappa^*(1-\beta)\omega^{1-\delta_\kappa} + \beta^2 c_\kappa^{**}\omega^{2(1-\delta_\kappa)} + o((1-\beta)\omega^{1-\delta_\kappa}) + o\left(\omega^{2(1-\delta_\kappa)}\right)}.$$

Consider the Fourier frequencies $\omega_j = 2\pi j/T$ for $j = 1, ..., n$ with $n = o(T)$. If $\nu > 1 - \delta_\kappa$, then for $j = 1, ..., n$, $(1-\beta) = o\left(\omega_j^{1-\delta_\kappa}\right)$ and

$$f_y(\omega_j) \underset{\omega_j \to 0^+}{\sim} \frac{1}{c_\kappa^{**}}\omega_j^{-2(1-\delta_\kappa)},$$

which also implies that $f_{\Delta y}(\omega_j) \underset{\omega_j \to 0^+}{\sim} \frac{1}{c_\kappa^{**}}\omega_j^{-2\delta_\kappa}$.

D.2 Autocorrelations

To derive our results we first need the following lemma whose proof is in a supplementary appendix.

Lemma 9 *Let f a spectral density with f, f' and f'' bounded, $f > 0$ in a neighborhood of the origin and $f'(0) = 0$. Let $|\lambda| \in (0,1)$ and $\omega_k = 2\pi k/T$, $k = o(T)$. Then,*

$$T^{\lambda-1}\sum_{j=1}^{T} j^{-\lambda} f(\omega_j)\cos(j\omega_k) = O\left(k^{\lambda-1}\right). \tag{41}$$

Now consider the autocovariances. The autocovariance function of y_t satisfies

$$\begin{aligned} \gamma_k &= \frac{1}{2\pi}\int_0^{2\pi} f_y(\omega) e^{ik\omega} d\omega & (42) \\ &= \frac{1}{2\pi}\int_0^{2\pi} f_y(\omega)\cos(k\omega)\, d\omega, & (43) \end{aligned}$$

to which the following finite sum converges (when it does converge)

$$\frac{1}{2\pi T}\sum_{j=1}^{T} f_y\left(\frac{2\pi j}{T}\right)\cos\frac{2\pi jk}{T} \underset{T\to\infty}{\to} \gamma_y(k). \tag{44}$$

In the proof, we use the theorems of Yong (1974) regarding sums of the type $\sum_{j=1}^{\infty} a_j \cos j\omega_k$ with $a_j \sim j^{-\alpha}$ for some $\alpha > 0$ as $\omega_k \to 0^+$. In the summation, we let the number of summation terms T and the evaluation value $\omega_k = 2\pi k T^{-1}$ tend to their limits, including $k \to \infty$. Zygmund (1935), section 1.23, shows that the sum (44) converges uniformly if $f_y(\omega_j)$ is of bounded variation and the latter follows from the results of section D.1. Our proof is therefore related to the method of Erdélyi (1956) considered by Lieberman and Phillips (2008), the difference is that we consider the limit of a finite sum (44) evaluated at Fourier frequencies where Lieberman and Phillips work with the integral representation (42).

We apply lemma 9 to expression (44) together with (40).When $\nu > 1-\delta_\kappa$, then $1-\beta = o(\omega_j)$ for all Fourier frequencies ω_j, $j = 1, ..., T$. Expression (40) hence implies that

$$\begin{aligned}
&\delta_\kappa \in (1/2,1) : f_y(\omega_j) \sim \frac{f_x(\omega_j)}{(2\pi)^{2(1-\delta_\kappa)} c_\kappa^{**} (j/T)^{2(1-\delta_\kappa)}};\\
&\delta_\kappa \in (0,1/2) : f_{\Delta y}(\omega_j) \sim \frac{f_x(\omega_j)}{(2\pi)^{-2\delta_\kappa} c_\kappa^{**} (j/T)^{-2\delta_\kappa}}.
\end{aligned} \tag{45}$$

We refer to lemma 9 where we let $\lambda = 2(1-\delta_\kappa)$ if $\delta_\kappa \in (1/2,1)$ and $\lambda = -2\delta_\kappa$ if $\delta_\kappa \in (0,1/2)$. Then for $k = o(T)$:

$$\delta_\kappa \in (1/2,1) : \gamma_y(k) = \begin{cases} O\left(k^{1-2\delta_\kappa}\right), & k \neq 0;\\ O(1), & k = 0. \end{cases}$$

$$\delta_\kappa \in (0,1/2) : \gamma_{\Delta y}(k) = \begin{cases} O\left(k^{-1-2\delta_\kappa}\right), & k \neq 0;\\ O(1), & k = 0. \end{cases}$$

E Proof of theorem 4

Consider the natural logarithm of spectral density $f_y(\omega)$ of y_t evaluated at the Fourier frequencies ω_j, for $j = 1, ..., n = o(T)$. Expression (40) implies that as $\omega_j \to 0^+$ and for $\nu > 1-\delta_\kappa$,

$$\begin{aligned}
&\delta_\kappa \in (1/2,1) : \log f_y(\omega_j) = \log f_x(\omega_j) - \log\left(\beta^2 c_\kappa^{**} \omega_j^{2(1-\delta_\kappa)} + o\left(\omega_j^{2(1-\delta_\kappa)}\right)\right),\\
&\delta_\kappa \in (0,1/2) : \log f_{\Delta y}(\omega_j) = \log f_x(\omega_j) - \log\left(\beta^2 c_\kappa^{**} \omega_j^{-2\delta_\kappa} + o\left(\omega_j^{-2\delta_\kappa}\right)\right).
\end{aligned}$$

We only consider the proof for the case where $\delta_\kappa \in (1/2, 1)$ as the proof for $\delta_\kappa \in (0, 1/2)$ follows the same lines. We denote by $h(\omega_j)$ the regressor that is used in the estimation, here $h(\omega_j) = -2\log\omega_j$. Hence, expression (40) implies that:

$$\begin{aligned}\log f_y(\omega_j) &= \log f_x(0) - \log\left(\beta^2 c_\kappa^{**}\right) + (1-\delta_\kappa) h(\omega_j) - \log(1 + o(1)) \\ &= \log f_x(0) - \log\left(\beta^2 c_\kappa^{**}\right) + (1-\delta_\kappa) h(\omega_j) + o(1).\end{aligned}$$

Now assume that f_y is estimated as $\widehat{f}_{y,T}$ and define $\phi_T(\omega_j) = \widehat{f}_{y,T}(\omega_j) / f_y(\omega_j)$. The ratio is defined since $f_y(\omega_j) > 0$ in a neighborhood of the origin, i.e. for T large enough. The estimator of the degree of memory, $\widehat{d}$, is the least squares estimator of the coefficient of $h(\omega_j)$ in the regression of $\log \widehat{f}_{y,T}(\omega_j)$ on a constant and $h(\omega_j)$,[18] where

$$\log \widehat{f}_{y,T}(\omega_j) = \log f_x(0) - \log\left(\beta^2 c_\kappa^{**}\right) + (1-\delta_\kappa) h(\omega_j) + \log\phi_T(\omega_j) + o_p(1).$$

Denoting by $\overline{\zeta}$ the average of $\zeta(\omega_j)$ over $j = 1, ..., n$ for any function ζ, the estimator satisfies

$$\widehat{d} = (1-\delta_\kappa) + \frac{1}{2}\frac{\sum_{j=1}^n \left(\log\phi_T(\omega_j) - \overline{\log\phi_T}\right)\left(h(\omega_j) - \overline{h}\right)}{\sum_{j=1}^n \left(h(\omega_j) - \overline{h}\right)^2} + o_p(1). \tag{46}$$

where as $n \to \infty$,

$$\sum_{j=1}^n \left(h(\omega_j) - \overline{h}\right)^2 \sim 4n. \tag{47}$$

We now make the high-level assumption that $\widehat{f}_{y,T}(\omega_j) \xrightarrow{p} f_y(\omega_j)$. The continuous mapping theorem implies that there exists $\tau_T \to \infty$, such that

$$\tau_T\left[\log \widehat{f}_{y,T}(\omega_j) - \log f_y(\omega_j)\right] \xrightarrow{p} 0, \tag{48}$$

i.e. $\tau_T \log\phi_T(\omega_j) \xrightarrow{p} 0$. Conditions for the consistency of the spectral density estimator can be found in various places in the literature and depend on the specific assumptions about x_t; see e.g. the references in the main text. It follows that $\sum_{j=1}^n \left(\log\phi_T(\omega_j) - \overline{\log\phi_T}\right)^2 = o_p\left(\frac{n}{\tau_T^2}\right)$ which, together with expression (47) and the Cauchy-Schwarz inequality, imply that $\widehat{d} - (1-\delta_\kappa) = o_p\left(\tau_T^{-1}\right) + o_p(1)$. The condition $\tau_T \to \infty$ as $T \to \infty$ is therefore sufficient to ensure that $\widehat{d} - (1-\delta_\kappa) \xrightarrow{p} 0$.

[18] The original Geweke and Porter-Hudak (1983) estimator used the periodogram for $\widehat{f}_{y,T}(\omega_j)$.

F Proof of theorem 5

Consider the partial sum of y_t, $S_T = \sum_{t=1}^{T} y_t = \sum_{t=1}^{T} (\beta a_t + x_t)$. Using expressions (1) and (7a), $a_t = \frac{1-g_t}{1-\beta g_t} a_{t-1} + \frac{g_t}{1-\beta g_t} x_t$ or

$$a_t = \left(\prod_{j=1}^{t} 1 - \frac{(1-\beta) g_j}{1-\beta g_j} \right) a_0 + \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} 1 - \frac{(1-\beta) g_j}{1-\beta g_j} \right) \frac{g_i x_i}{1-\beta g_i},$$

with $\prod_{j=t+1}^{t} \left(1 - \frac{(1-\beta)g_j}{1-\beta g_j}\right) \equiv 1$. When $g_i \to 0$, $\frac{g_i}{1-\beta g_i} = g_i + o(g_i)$, so the order of magnitude of a_t is the same as that of[19]

$$a_t^* = \left(\prod_{j=1}^{t} 1 - (1-\beta) g_j \right) a_0 + \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} 1 - (1-\beta) g_j \right) g_i x_i. \tag{49}$$

Hence, we can infer the order of magnitude of $var(S_T)$ from that of $var(S_T^*)$, where $S_T^* = \sum_{t=1}^{T} (\beta a_t^* + x_t)$. Using (49), S_T^* can be written as

$$S_T^* = \beta h_{T+1} a_0 + \sum_{t=1}^{T} \phi_{T,t} x_t,$$

where $\phi_{T,t} = 1 + \beta g_t \sum_{i=t}^{T} \prod_{j=t+1}^{i} (1 - (1-\beta) g_j)$ and $h_t = \sum_{i=1}^{t-1} \prod_{j=1}^{i} (1 - (1-\beta) g_j)$. Note that $\phi_{T,t} = 1 + \beta \frac{g_t}{k_t} (h_{T+1} - h_t)$, where $k_t = \prod_{j=1}^{t} (1 - (1-\beta) g_j)$.

For clarity, we first consider the case when x_t is serially uncorrelated, and treat the general case at the end. The variance of S_T^* is given by

$$var[S_T^*] = \beta^2 h_{T+1}^2 var(a_0) + \sigma_x^2 \sum_{t=1}^{T} \phi_{T,t}^2,$$

where $\sigma_x^2 = var[x_t]$. We study each of the two terms on the right hand side of the above expression.

The asymptotic rates of h_t and k_t depend on the value of β. Since $g_i \sim i^{-1}$, $g_i^2 = o(g_i)$. This implies that $\log(1 - (1-\beta) g_i) = -(1-\beta) g_i + o(g_i)$ and $\log k_t = -(1-\beta) \log t + o(\log t)$. Thus, $g_t / k_t \sim t^{-1} / t^{-1+\beta} = t^{-\beta}$. Turning to $h_t = \sum_{i=1}^{t-1} k_i$,

$$h_t \sim \begin{cases} \beta^{-1} t^{\beta} + o\left(t^{\beta}\right), & \text{if } 0 < \beta \leq 1; \\ \log t + o(\log t), & \text{if } \beta = 0; \\ \zeta(1-\beta) + o(1), & \text{if } \beta < 0, \end{cases} \tag{50}$$

[19] In the specific situation where $g_1 = 1$ the impact of a_0 on a_t is zero contrary to that on a_t^*. This only concerns g_1 since $g_{i+1} < g_i \leq 1$ for all $i \geq 1$; it does not affect the magnitude of $var(S_T)$ as we show later.

where $\zeta(u)$ is Riemann's zeta function evaluated at $u > 1$. (the case $\beta = 0$ is included for completeness, since it plays no role in the asymptotic rates of $var(S_T^*)$). It follows that as $t \to \infty$, for $t \leq T$,

$$\phi_{T,t} = O\left(1 + \left(\frac{T}{t}\right)^{\beta}\right), \text{ for } \beta \leq 1. \tag{51}$$

So the two contributions to the variance of S_T^* are:

$$\sum_{t=1}^{T} \phi_{T,t}^2 = \begin{cases} O\left(T^{2\beta}\right), & \text{if } 1/2 < \beta \leq 1; \\ O\left(T \log T\right), & \text{if } \beta = 1/2; \\ O\left(T\right), & \text{if } \beta < 1/2, \end{cases} \tag{52}$$

and, corresponding to the impact of a_0 :

$$h_{T+1}^2 = \begin{cases} O\left(T^{2\beta}\right), & \text{if } 0 < \beta \leq 1; \\ O\left(\left(\log T\right)^2\right), & \text{if } \beta = 0; \\ O\left(1\right), & \text{if } \beta < 0. \end{cases}$$

The latter expression shows that if $a_0 = O_p(1)$, its contribution to $T^{-1}Var[S_T^*]$ vanishes asymptotically when $\beta \leq 1/2$. The result of the theorem then follows from the rates in (52).

Now, we turn to the general case where x_t is not serially uncorrelated, and denote by $\gamma_x(\cdot)$ its autocovariance function. Then $var(S_T^*)$ contains the additional term

$$2 \sum_{t=1}^{T-1} \phi_{T,t} \sum_{i=1}^{T-t} \phi_{T,t+i} \gamma_x(i). \tag{53}$$

Since the autocovariance function of x_t decays exponentially, there exists $\theta \in (0,1)$ such that $|\gamma_x(i)| \leq \sigma_x^2 \theta^i$ for all i. Then,

$$\left| \sum_{t=1}^{T-1} \phi_{T,t} \sum_{i=1}^{T-t} \phi_{T,t+i} \gamma_x(i) \right| \leq \sigma_x^2 \sum_{t=1}^{T-1} \phi_{T,t} \sum_{i=1}^{T-t} \phi_{T,t+i} \theta^i.$$

We will show that the term $\sum_{i=1}^{T-t} \phi_{T,t+i} \theta^i$ is of the same order of magnitude as $\phi_{T,t}$, given in expression (51), which suffices to establish the result of the theorem in the general case.

Expression (51) implies that $\phi_{T,t+i}\theta^i = O\left(\theta^i + T^{\beta}\theta^i \left(\frac{1}{t+i}\right)^{\beta}\right)$ and so

$$\sum_{i=1}^{T-t} \phi_{T,t+i} \theta^i = O\left(1 + T^{\beta} \sum_{i=1}^{T-t} \frac{\theta^i}{(t+i)^{\beta}}\right). \tag{54}$$

For $\beta \leq 0$, it is clear that $\sum_{i=1}^{T-t} \phi_{T,t+i}\theta^i = O(1)$ since $\sum_{i=1}^{T-t} \theta^i (t+i)^{-\beta} = O(1)$. For $\beta > 0$, we show that $\sum_{i=1}^{T-t} \frac{\theta^i}{(t+i)^\beta} = \theta^{-t} \sum_{i=t+1}^{T} \frac{\theta^i}{i^\beta} = O\left(t^{-\beta}\right)$, which implies that $\sum_{i=1}^{T-t} \phi_{T,t+i}\theta^i = O\left(\left(\frac{T}{t}\right)^\beta\right)$ from (54).

We prove that $\sum_{i=1}^{T-t} \frac{\theta^i}{(t+i)^\beta} = O\left(t^{-\beta}\right)$ using the Lerch Transcendent, see Erdélyi et al. (1953), pp. 27-31, which is defined as

$$\Phi(z, s, \alpha) = \sum_{n=0}^{\infty} \frac{z^n}{(n+\alpha)^s},$$

and is finite for $z < 1$, $s > 0$, $\alpha > 0$ or $z = 1$, $s > 1$, $\alpha > 0$. From

$$\sum_{n=1}^{t} \frac{z^n}{(n+\alpha)^s} = z\Phi(z, s, \alpha+1) - z^{t+1}\Phi(z, s, \alpha+t+1),$$

it follows that

$$\sum_{i=1}^{T-t} \frac{\theta^i}{(t+i)^\beta} = \theta\Phi(\theta, \beta, t+1) - \theta^{T-t+1}\Phi(\theta, \beta, T+1),$$

and we notice that, for $t > 1$:

$$t^\beta \Phi(\theta, \beta, t) = \sum_{i=0}^{\infty} \frac{\theta^i}{(i/t+1)^\beta} \in \left(\sum_{i=0}^{\infty} \frac{\theta^i}{(i+1)^s}, \sum_{i=0}^{\infty} \theta^i\right) = \left(\Phi(\theta, \beta, 1), \frac{1}{1-\theta}\right).$$

Hence, as $t \to \infty$,

$$\Phi(\theta, \beta, t) = O\left(t^{-\beta}\right), \tag{55}$$

which gives the desired result as $t \to \infty$ and for $t \leq T$:

$$\sum_{i=1}^{T-t} \frac{\theta^i}{(t+i)^\beta} = O\left(t^{-\beta}\right) + O\left(\theta^{T-t+1}T^{-\beta}\right) = O\left(t^{-\beta}\right).$$

G Proof of expression (17)

Under CGLS learning the algorithm is

$$\kappa_t(L) = \bar{g} \sum_{j=0}^{t-1} (1-\bar{g})^j L^j,$$

and $\zeta_t = a_0 (1-\bar{g})^t$. Hence

$$\begin{aligned} m(\kappa_t) &= \bar{g} \sum_{j=1}^{t-1} j (1-\bar{g})^j = -\bar{g}(1-\bar{g}) \frac{\partial}{\partial \bar{g}} \sum_{j=0}^{t-1} (1-\bar{g})^j \\ &= (1-\bar{g}) \frac{1 - (1-\bar{g})^{t-1}[1 + (t-1)\bar{g}]}{\bar{g}}. \end{aligned}$$

Now consider $m(\kappa_T)$, and assume that $\bar{g} = c_g T^{-\lambda}$. Then $(1-\bar{g})^{T-1} = \exp\left\{(T-1)\log\left(1-c_g T^{-\lambda}\right)\right\}$ and as $T \to \infty$

$$(1-\bar{g})^{T-1} \sim \exp\left\{-c_g \frac{T-1}{T^{\lambda}}\right\} \to \begin{cases} 0, & \text{if } \lambda < 1; \\ e^{-c_g}, & \text{if } \lambda = 1. \end{cases}$$

Turning to the mean lag, for $\lambda < 1$ $(1-\bar{g})^{t-1}[1+(t-1)\bar{g}] \to 0$ so $m(\kappa_T) \sim \frac{T^{\lambda}}{c_g}$. When $\lambda = 1$, $m(\kappa_T) \sim \frac{1-e^{-c_g}[1+c_g]}{c_g} T$, which proves (17).

H Proof of theorem 6

Under the stated assumptions, the estimator a_t is generated by

$$a_t = \frac{\bar{g}}{1-\beta\bar{g}} \sum_{i=1}^{t} \left(1 - \frac{(1-\beta)\bar{g}}{1-\beta\bar{g}}\right)^{t-i} x_i.$$

When β is local to unity and $\bar{g}$ local to zero, the magnitude of a_t is the same as that of

$$a_t^* = \bar{g} \sum_{i=1}^{t} (1-(1-\beta)\bar{g})^{t-i} x_i,$$

which is simpler to analyze using existing results. We will obtain the order of magnitude of $S_T^* = \sum_{t=1}^{T} (\beta a_t^* + x_t)$, which is the same as that of $\sum y_t$.

Define $\xi_t = \bar{g}^{-1} a_t^*$ such that

$$\xi_t = \sum_{i=1}^{t} (1-(1-\beta)\bar{g})^{t-i} x_i,$$

with $(\beta, \bar{g}) = \left(1 - c_\beta T^{-\nu}, c_g T^{-\lambda}\right)$ for $(\nu, \lambda) \in [0,1]^2$. Several cases arise depending on the values of λ, ν. These correspond to a_t exhibiting an exact unit root for $\bar{g} = 0$ or $\beta = 1$, a near-unit root for $\lambda + \nu = 1$ (see Chan and Wei, 1987, and Phillips 1987), a moderate-unit root for $\lambda + \nu \in (0,1)$ (see Giraitis and Phillips, 2006, Phillips and Magdalinos, 2007 and Phillips, Magdalinos and Giraitis, 2010) and a very-near-unit root for $\lambda + \nu > 1$ (see Andrews and Guggenberger, 2007). Their results imply:

$$\xi_T = \begin{cases} O_p(1), & \lambda = \nu = 0; \\ O_p\left(T^{(\lambda+\nu)/2}\right), & \lambda + \nu \in (0,1); \\ O_p\left(T^{1/2}\right), & \lambda + \nu \geq 1. \end{cases}$$

To derive the magnitude of $S_T^* = \beta\bar{g}\sum_{t=1}^T \xi_{t-1} + \sum_{t=1}^T x_t$ we notice that:

$$\sum_{t=1}^{T}\xi_t = \sum_{t=1}^{T}\sum_{i=1}^{t}\left(1-\left(1-\beta\right)\bar{g}\right)^{t-i}x_i = \sum_{t=1}^{T}\frac{1-\left(1-\left(1-\beta\right)\bar{g}\right)^{T-t+1}}{1-\left(1-\left(1-\beta\right)\bar{g}\right)}x_t,$$

i.e.

$$\sum_{t=1}^{T}\xi_t = \frac{1}{\left(1-\beta\right)\bar{g}}\left[\sum_{t=1}^{T}x_t - \left(1-\left(1-\beta\right)\bar{g}\right)\xi_T\right].$$

Hence

$$\bar{g}\sum_{t=1}^{T}\xi_t = \frac{1}{\left(1-\beta\right)}\left(\sum_{t=1}^{T}x_t - \xi_T\right) + \bar{g}\xi_T. \tag{56}$$

We start with the case $\nu+\lambda<1$, where $\xi_T = o\left(\sum_{t=1}^T x_t\right)$. Expression (56) implies that $\bar{g}\sum_{t=1}^T\xi_t = O_p\left(T^{1/2+\nu}\right)$ and hence

$$\mathsf{sd}\left(T^{-1/2}S_T^*\right) = O\left(T^{\nu}\right).$$

If $\nu+\lambda=1$, then Phillips (1987) shows that

$$T^{-1/2}\left(\sum_{t=1}^{T}x_t - \xi_T\right) = T^{-1/2}\sum_{i=1}^{T}\left(1-\left(1-\left(1-\beta\right)\bar{g}\right)^{T-i}\right)x_i$$
$$\Rightarrow \int_0^1\left(1-e^{-c_\beta c_g(1-r)}\right)dW\left(r\right) = O_p\left(1\right),$$

where $T^{-1/2}\sum_{t=1}^{\lceil rT\rceil}x_t \Rightarrow W\left(r\right)$, where $W\left(\cdot\right)$ is a Brownian motion and $\Rightarrow$ denotes weak convergence of the associated probability measure. It follows that $\sum_{t=1}^T x_t - \xi_T = O\left(T^{1/2}\right)$ and expression (56) implies that $\bar{g}\sum_{t=1}^T\xi_t = O_p\left(T^{1/2+\nu}\right)$. Hence

$$\mathsf{sd}\left(T^{-1/2}S_T^*\right) = O\left(T^{\nu}\right) = O\left(T^{1-\lambda}\right).$$

Now, if $\nu+\lambda>1$,

$$\sum_{t=1}^{T}x_t - \xi_T = \sum_{i=0}^{T-1}\left[1-\left(1-\left(1-\beta\right)\bar{g}\right)^{i}\right]x_{T-i}$$
$$= \left(\left(1-\beta\right)\bar{g}\right)\sum_{i=0}^{T-1}\left[i+O\left(i^2\left(\left(1-\beta\right)\bar{g}\right)\right)\right]x_{T-i}.$$

It is well known that $\sum_{i=0}^{T-1} i x_{T-i} = O_p\left(T^{3/2}\right)$ and $\sum_{i=0}^{T-1} i^2 x_{T-i} = O_p\left(T^{5/2}\right)$ (see e.g. Hamilton 1994, chap. 17). Hence $(1-\beta)\,\bar{g}\sum_{i=0}^{T-1} i^2 x_{T-i} = o\left(\sum_{i=0}^{T-1} i x_{T-i}\right)$, and, in expression (56):

$$\frac{1}{(1-\beta)}\left(\sum_{t=1}^{T} x_t - \xi_T\right) + \bar{g}\xi_T = O_p\left(T^{3/2-\lambda}\right) + O_p\left(T^{1/2-\lambda}\right).$$

When $\lambda < 1$, $3/2 - \lambda > 1/2$ so $\sum_{t=1}^{T} x_t = o_p\left(\bar{g}\sum_{t=1}^{T}\xi_{t-1}\right)$, and the order of magnitude of S_T^* follows from that of $\bar{g}\sum_{t=1}^{T}\xi_{t-1}$:

$$\mathsf{sd}\left(T^{-1/2}S_T^*\right) = O\left(T^{1-\lambda}\right).$$

If $\lambda = 1$, $\sum_{t=1}^{T} x_t = O_p\left(\bar{g}\sum_{t=1}^{T}\xi_{t-1}\right)$ and the previous expression also applies.

I Derivation of expression (20)

We show that expression (19) can be written as (20). We start from expression (19):

$$Y_t = (1-\beta)\sum_{j=0}^{\infty}\beta^j E_t\left(\gamma_1' z_{t+j}\right) + \beta\sum_{j=0}^{\infty}\beta^j E_t\left(\gamma_2' z_{t+j}\right),$$

and notice the identity $(1-\beta)\sum_{j=0}^{\infty}\beta^j z_{t+j} = z_t + \sum_{j=1}^{\infty}\beta^j \Delta z_{t+j}$, so

$$Y_t - \left(\gamma_1' + \frac{\beta}{1-\beta}\gamma_2'\right) z_t = \frac{\beta}{1-\beta} E_t \Delta Y_{t+1}, \tag{57}$$

where we used

$$\Delta Y_{t+1} = \frac{1-\beta}{\beta}\sum_{j=1}^{\infty}\beta^j E_{t+1}\left(\gamma_1' + \frac{\beta}{1-\beta}\gamma_2'\right)\Delta z_{t+j}.$$

Differencing (57) yields

$$\Delta Y_t - \left(\gamma_1' + \frac{\beta}{1-\beta}\gamma_2'\right)\Delta z_t = \frac{\beta}{1-\beta}\left(E_t\Delta Y_{t+1} - E_{t-1}\Delta Y_t\right).$$

Re-arranging yields:

$$(1-\beta)\,\Delta Y_t - (1-\beta)\left(\gamma_1' + \frac{\beta}{1-\beta}\gamma_2'\right)\Delta z_t = \beta\left(E_t\Delta Y_{t+1} - E_{t-1}\Delta Y_t\right),$$

or

$$\Delta Y_t - \beta Y_t + \beta E_{t-1} Y_t - \left((1-\beta)\,\gamma_1' + \beta\gamma_2'\right)\Delta z_t = \beta E_t \Delta Y_{t+1},$$

or

$$\Delta Y_t = \beta E_t \Delta Y_{t+1} + \left((1-\beta)\,\gamma_1^{'} + \beta\gamma_2^{'} \right) \Delta z_t + \beta r_t,$$

where we used the definition of the innovation $r_t = Y_t - E_{t-1}Y_t$. This can be written as (19) by defining $x_t = \left((1-\beta)\,\gamma_1^{'} + \beta\gamma_2^{'} \right) \Delta z_t + \beta r_t$, $y_t = \Delta Y_t$ and $y_{t+1}^e = E_t \Delta Y_{t+1}$.

References

Abadir, K. and G. Talmain (2002). Aggregation, persistence and volatility in a macro model. *Review of Economic Studies 69*(4), 749–79.

Andrews, D. W. K. and P. Guggenberger (2007). Asymptotics for stationary very nearly unit root processes. *Journal of Time Series Analysis 29*(1), 203–212.

Andrews, D. W. K. and D. Pollard (1994). An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review / Revue Internationale de Statistique 62*(1), pp. 119–132.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics 73*, 5–59.

Baillie, R. T. and T. Bollerslev (1994a). Cointegration, fractional cointegration, and exchange rate dynamics. *Journal of Finance 49*, 737–745.

Baillie, R. T. and T. Bollerslev (1994b). The long memory of the forward premium. *Journal of International Money and Finance 13*, 565–571.

Baillie, R. T. and T. Bollerslev (2000). The forward premium anomaly is not as bad as you think. *Journal of International Money and Finance 19*, 471488.

Beran, J. (1989). A test of location for data with slowly decaying serial correlations. *Biometrika 76*(2), pp. 261–269.

Beran, J. (1994). *Statistics for Long-Memory Processes*. Chapman & Hall.

Berenguer-Rio, V. and J. Gonzalo (2011). Summability of stochastic processes (a generalization of integration and co-integration valid for non-linear processes). Working paper, Universidad Carlos 3.

Bobkoski, M. (1983). Hypothesis testing in nonstationary time series. Unpublished PhD thesis, Dept. of Statistics, University of Wisconsin, Madison.

Branch, W. and G. Evans (2010). Asset return dynamics and learning. *Review of Financial Studies 23*, 1651–80.

Brillinger, D. R. (1975). *Time Series Data Analysis and Theory*. New York: Holt, Rinehart and Winston. Reprinted in 2001 as a SIAM Classic in Applied Mathematics.

Brock, W. A. and C. H. Hommes (1997). A rational route to randomness. *Econometrica 65*, 1059–95.

Cagan, P. (1956). The monetary dynamics of hyper-inflation. In M. Friedman (Ed.), *Studies in the Quantity Theory of Money*, pp. 25–120. Chicago: University of Chicago Press.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1996). *The Econometrics of Financial Markets*. London: Princeton University Press.

Campbell, J. Y. and N. G. Mankiw (1987). Are output fluctuations transitory? *Quarterly Journal of Economics 102*(4), 857–880.

Campbell, J. Y. and R. J. Shiller (1987). Cointegration and tests of present value models. *Journal of Political Economy 95*, 1062–1088.

Campbell, J. Y. and R. J. Shiller (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies 1*(3), 195–228.

Campbell, J. Y. and M. Yogo (2006). Efficient tests of stock return predictability. *Journal of Financial Economics 81*(1), 27–60.

Cavanagh, C. (1985). Roots local to unity. Unpublished manuscript, Harvard University.

Chakraborty, A. and G. Evans (2008). Can perpetual learning explain the forward premium puzzle? *Journal of Monetary Economics 55*, 477–90.

Chambers, M. J. (1998). Long memory and aggregation in macroeconomic time series. *International Economic Review 39*(4), pp. 1053–1072.

Chan, N. H. and C. Z. Wei (1987). Asymptotic inference for nearly nonstationary AR(1) processes. *Annals of Statistics 15*(3), 1050–1063.

Cheung, Y.-W. (1993). Long-memory in foreign exchange rates. *Journal of Business and Economic Statistics 11*(1), 93–101.

Cheung, Y.-W. and K. S. Lai (1993). A fractional cointegration analysis of purchasing power parity. *Journal of Business and Economic Statistics 11*(1), 103–112.

Chevillon, G., M. Massmann, and S. Mavroeidis (2010). Inference in models with adaptive learning. *Journal of Monetary Economics 57*(3), 341–51.

Clarida, R., J. Galí, and M. Gertler (1999). The Science of Monetary Policy: A New Keynesian Perspective. *Journal of Economic Literature 37*(4), 1661–1707.

Cogley, T. (2002). A simple adaptive measure of core inflation. *Journal of Money, Credit and Banking 34*(1), 94–113.

Davidson, J. and N. Hashimzade (2008). Alternative frequency and time domain versions of fractional Brownian motion. *Econometric Theory 24*(1), 256–293.

Davidson, J. and N. Hashimzade (2009). Type I and type II fractional Brownian motions: a reconsideration. *Computational Statistics and Data Analysis 53*(6), 2089–2106.

Davidson, J. and P. Sibbertsen (2005). Generating schemes for long memory processes: regimes, aggregation and linearity. *Journal of Econometrics 128*(2), 253–82.

Davidson, J. and T. Teräsvirta (2002). Long memory and nonlinear time series. *Journal of Econometrics 110*(2), 105–12.

Delgado, M. A. and P. M. Robinson (1996). Optimal spectral kernel for long-range dependent time series. *Statistics and Probability Letters 30*, 37–43.

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

Diebold, F. X. and G. D. Rudebusch (1991). On the power of dickey-fuller tests against fractional alternatives. *Economics Letters 35*(2), 155 – 160.

Durbin, J. and S. J. Koopman (2008). *Time Series Analysis by State Space Methods*. Oxford University Press. 2nd ed.

Engel, C. and K. D. West (2005). Exchange rates and fundamentals. *Journal of Political Economy 113*(3), 485–517.

Erdélyi, A. (1956). *Asymptotic Expansions*. USA: Dover Publication Inc.

Erdélyi, A., W. Magnus, F. Oberhettinger, and F. G. Tricomi (1953). *Higher Transcendental Functions. Vol. I*. New York-Toronto-London: McGraw-Hill Book Company, Inc.

Evans, G. W. and S. Honkapohja (2001). *Learning and Expectations in Macroeconomics*. Princeton: Princeton University Press.

Evans, G. W. and S. Honkapohja (2009). Expectations, learning and monetary policy: An overview of recent rersearch. In K. Schmidt-Hebbel and C. Walsh (Eds.), *Monetary Policy under Uncertainty and Learning*, pp. 27–76. Santiago: Central Bank of Chile.

Fama, E. F. (1984). Forward and spot exchange rates. *Journal of Monetary Economics 14*(3), 319–338.

Fama, E. F. and G. W. Schwert (1977). Asset returns and inflation. *Journal of Financial Economics 5*, 115–46.

Frankel, J. A. and K. A. Froot (1987). Using survey data to test standard propositions regarding exchange rate expectations. *American Economic Review 77*(1), 133–153.

Geweke, J. and S. Porter-Hudak (1983). he estimation and application of long memory time series models. *Journal of Time Series Analysis 4*, 221–238.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*, 1545–1578.

Giraitis, L. and P. C. B. Phillips (2006). Uniform limit theory for stationary autoregression. *Journal of Time Series Analysis 27*, 51–60.

Gonzalo, J. and J.-Y. Pitarakis (2006). Threshold effects in cointegrating relationships. *Oxford Bulletin of Economics and Statistics 68*, 813–833.

Gourieroux, C., J. J. Laffont, and A. Monfort (1982). Rational expectations in dynamic linear models: Analysis of the solutions. *Econometrica 50*(2), pp. 409–425.

Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics 14*(2), 227–238.

Granger, C. W. J. and Z. Ding (1996). Varieties of long memory models. *Journal of econometrics 73*(1), 61–77.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Heyde, C. C. and Y. Yang (1997). On defining long range dependence. *Journal of Applied Probability 34*, 939–944.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers 116*, 770–99.

Hurvich, C. M., R. Deo, and J. Brodsky (1998). The mean squared error of the Geweke and Porter-Hudak's estimator of the long memory parameter of a long-memory time series. *Journal of Time Series Analysis 19*(1), 19–46.

Kim, C. and P. Phillips (1999). Log periodogram regression: the nonstationary case. Cowles foundation discussion paper, Yale University.

Lieberman, O. and P. C. Phillips (2008). A complete asymptotic series for the autocovariance function of a long memory process. *Journal of Econometrics 147*(1), 99 – 103.

Lo, A. W. (1991). Long-term memory in stock market prices. *Econometrica 59*, 1279–313.

Malmendier, U. and S. Nagel (2011). Depression babies: Do macroeconomic experiences affect risk-taking? *Quarterly Journal of Economics 126*(1), 373–416.

Marinucci, D. and P. Robinson (1999). Alternative forms of fractional brownian motion. *Journal of Statistical Planning and Inference 80*(1), 111–122.

Marinucci, D. and P. M. Robinson (2001). Semiparametric fractional cointegration analysis. *Journal of Econometrics 105*, 225–47.

Maynard, A. and P. C. B. Phillips (2001). Rethinking an old empirical puzzle: Econometric evidence on the forward discount anomaly. *Journal of Applied Econometrics 16*(6), 671–708.

Milani, F. (2007). Expectations, learning and macroeconomic persistence. *Journal of Monetary Economics 54*(7), 2065–2082.

Miller, J. I. and J. Y. Park (2010). Nonlinearity, nonstationarity, and thick tails: How they interact to generate persistence in memory. *Journal of Econometrics 155*(1), 83 – 89.

Müller, U. and M. W. Watson (2008). Testing models of low-frequency variability. *Econometrica 76*(5), 9791016.

Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association 55*(290), 229–306.

Nerlove, M. (1958). Adaptive expectations and cobweb phenomena. *The Quarterly Journal of Economics 72*(2), 227–240.

Parke, W. R. (1999). What is fractional integration? *Review of Economics and Statistics 81*(4), 632–638.

Perron, P. and Z. Qu (2007). An analytical evaluation of the log-periodogram estimate in the presence of level shifts. working paper, Boston University.

Phillips, P. C. B. (1987). Towards a unified asymptotic theory for autoregression. *Biometrika 74*(3), 535–547.

Phillips, P. C. B. (2007). Unit root log periodogram regression. *Journal of Econometrics 138*(1), 104–24.

Phillips, P. C. B. and T. Magdalinos (2007). Limit theory for moderate deviations from a unit root. *Journal of Econometrics 136*, 115–130.

Phillips, P. C. B., T. Magdalinos, and L. Giraitis (2010). Smoothing local-to-moderate unit root theory. *Journal of Econometrics 158*(2), 274–79.

Phillips, P. C. B. and V. Solo (1992). Asymptotics for linear processes. *Annals of Statistics 20*(2), 971–1001.

Robinson, P. M. (1994a). Rates of convergence and optimal spectral bandwidth for long range dependence. *Probability Theory and Related Fields 99*, 443–473.

Robinson, P. M. (1994b). Semiparametric analysis of long-memory time series. *Annals of Statistics 22*, 515–39.

Robinson, P. M. (1995a). Gaussian semiparametric estimation of long range dependence. *Annals of Statistics 23*, 1630–61.

Robinson, P. M. (1995b). Log-periodogram regression of time series with long range dependence. *Annals of Statistics 23*, 1048–72.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. 42*(1), 43–47.

Rozeff, M. S. (1984). Dividend yields are equity risk premiums. *Journal of Portfolio Management 11*, 68–75.

Sargent, T. J. (1993). *Bounded Rationality in Macroeconomics*. Oxford University Press.

Sargent, T. J. (1999). *The conquest of American inflation*. USA: Princeton University Press.

Shiller, R. J. (1989). Comovements in stock prices and comovements in dividends. *Journal of Finance 44*, 719–29.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics 54*, 375–421.

Stock, J. H. (1994). Unit roots, structural breaks and trends. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, Chapter 46, pp. 2739–2841. Elsevier.

Tanaka, K. (1999). The nonstationary fractional unit root. *Econometric Theory 15*, 549–82.

Teverovsky, V. and M. Taqqu (1997). Testing for long-range dependence in the presence of shifting means or a slowly declining trend, using a variance-type estimator. *Journal of Time Series Analysis 18*, 279–304.

White, H. (2000). *Asymptotic Theory for Econometricians*. Academic Press Inc. 2nd ed.

Yong, C. H. (1974). *Asymptotic Behaviour of Trigonometric Series*. Chinese University of Hong Kong.

Zaffaroni, P. (2004). Contemporaneous aggregation of linear dynamic models in large economies. *Journal of Econometrics 120*(1), 75 – 102.

Zygmund, A. (1935). *Trigonometrical Series*. Warsaw: Monografje Matematyczne V.

ESSEC Business School
Avenue Bernard Hirsch
BP 50105
95021 Cergy-Pontoise Cedex
France

Tél. +33 (0)1 34 43 30 00
Fax +33 (0)1 34 43 30 01

www.essec.fr

ESSEC Executive Education
CNIT BP 230
92053 Paris-La Défense
France

Tél. +33 (0)1 46 92 49 00
Fax +33 (0)1 46 92 49 90

http://formation.essec.fr

ESSEC Business School
Singapore Campus
100 Victoria Street
National Library Building # 13-02
Singapore 188064

essecasia@essec.fr
Tél. +65 6884 9780
Fax +65 6884 9781

www.essec.edu

Informations
Alison Bougi
+33 (0)1 34 43 33 58
bougi@essec.edu
www.essec.fr
research.center@essec.fr

ISSN 1291-9616

ESSEC BUSINESS SCHOOL

CCI À vos côtés, précisément. AACSB EFMD EQUIS ACCREDITED