Perpetual Learning and Apparent Long Memory*

Guillaume Chevillon† ESSEC Business School

Sophocles Mavroeidis‡ University of Oxford

November 1, 2017

Abstract

This paper studies the low frequency dynamics in forward looking models where expectations are formed using perpetual learning such as constant gain least squares. We show that if the coefficient on expectations is sufficiently close to unity, perpetual learning induces strong persistence that is empirically indistinguishable from long memory. We apply this result to present value models of stock prices and exchange rates and find that perpetual learning can explain the long memory observed in the data.

JEL Codes: C1, E3;

Keywords: Long Memory, Consistent Expectations, Perpetual Learning, Present-Value Models.

*Chevillon acknowledges research support from Labex MME-DII and CREST. Mavroeidis would like to thank the European Commission for research support under a FP7 Marie Curie Fellowship CIG 293675.

†ESSEC Business School, Department of Information Systems, Decision Sciences and Statistics, Avenue Bernard Hirsch, BP50105, 95021 Cergy-Pontoise cedex, France. Email: guillaume.chevillon@essec.edu.

‡Department of Economics and INET at Oxford, University of Oxford, Manor Road, Oxford, OX1 3UQ, United Kingdom. Email: sophocles.mavroeidis@gmail.com.

1 Introduction

An extensive literature has studied persistence in models where agents form their expectations using adaptive learning (e.g., Bullard and Eusepi, 2005, Milani, 2007, Carceles-Poveda and Giannitsarou, 2008, Eusepi and Preston, 2011, Slobodyan and Wouters, 2012). Most of the literature has focused on persistence at business cycle frequencies, yet some authors have studied persistence at lower frequencies when the coefficient measuring the feedback that expectations have on the endogenous variables, say β, is close to unity. In the recent literature, Bullard, Evans and Honkapohja (2010), Branch and Evans (2011, 2017), Evans, Honkapohja, Sargent and Williams (2013), have all shown that as β gets closer to unity, near unit root behaviors (near random walks) can either be generated or rationalized. Chevillon and Mavroeidis (2017, henceforth CM) have further characterized the type of low frequency persistence that can be generated. They showed in particular that decreasing gain learning in forward looking models can generate strong persistence at low frequencies, akin to long memory.

This paper contributes to the above literature by extending the analysis in CM to the case of perpetual learning, which is the term typically used to refer to constant gain least-squares learning, see, e.g., Orphanides and Williams (2004). This extension is important because perpetual learning is more common in applied work and has different dynamic properties than decreasing gain learning (see Evans and Honkapohja, 2001), so the results in CM do not necessarily carry over to perpetual learning.

Under adaptive learning, agents form expectations by estimating models on real-time data. Their estimates and expectations can be characterized by stochastic recursive algorithms, and they can be divided into two categories: decreasing gain and constant gain algorithms. The so-called "gain" is a deterministic sequence that governs how responsive estimate revisions are to new data. The prototypical example of a decreasing gain algorithm is recursive least squares, according to which every new data point carries the same weight as all past data. In contrast, under constant gain algorithms past data gets discounted, which is desirable when agents expect the parameters of their forecasting model to change over time. Under constant gain learning, agents' beliefs do not converge, so learning dynamics remain in perpetuity.

We find that, in contrast to decreasing gain learning, perpetual learning does not generate long memory in forward-looking models. However, we show that if the coefficient on expectations is sufficiently large, i.e., if the feedback from expectations to the variables in self-referential models is sufficiently strong, the dynamics generated by perpetual learning can be empirically indistinguishable from long memory. Hence, perpetual learning can generate *apparent* long memory. An interesting by-product of the analysis is that agents can learn to believe in long memory in the sense that a perceived law of motion with long memory is

shown to be self-confirming.

The above theoretical results are used to shed some light on two well-known empirical puzzles in macroeconomics and finance. Specifically, we study the Campbell and Shiller (1987) model for stock prices, and the models of Engel and West (2005) for exchange rates. Under rational expectations, both models exhibit features that appear counterfactual and have led to the famous empirical puzzles of excess return predictability and the forward premium anomaly. Some explanations for these puzzles that have been proposed in the literature rely on the presence of long memory that is attributed to persistent shocks and is therefore of exogenous origin, see Baillie and Bollerslev (2000), Maynard and Phillips (2001) and Schotman et al. (2008). Here, we examine whether learning can account for the persistence at low frequencies observed in the data even when the exogenous shocks have short memory, and we find that it can.

The paper is organized as follows. In Section 2, we set up the model and assumptions. We follow CM and restrict our attention to representative agent linear models to avoid introducing any form of long memory which is not directly caused by learning, such as aggregation, nonlinearity or structural breaks. Sections 3 presents our analytical results for two classes of perpetual learning algorithms. Simulations follow in Section 4. Section 5 provides the empirical applications to two present value models. Section 6 concludes. Proofs are given in the Appendix. A Supplementary Appendix, available online, contains proofs of auxiliary lemmas.

Throughout the paper, $f(x) \sim g(x)$ as $x \to a$ means $\lim_{x\to a} f(x)/g(x) = 1$; $O(\cdot)$ and $o(\cdot)$ denote standard orders of magnitude; and $f(x) \asymp g(x)$ means "exact rate", *i.e.*, $f(x) = O(g(x))$ and $g(x) = O(f(x))$. Corresponding magnitudes in probability are denoted by $O_p(\cdot)$, $o_p(\cdot)$ and $\overset{p}{\asymp}$. Also, we use the notation $\mathsf{sd}(X)$ to refer to the standard deviation $\sqrt{\mathsf{Var}(X)}$.

2 Definitions and assumptions

Consider the following forward-looking model that links an endogenous variable y_t to an exogenous process x_t:

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

where y_{t+1}^e denotes the expectation of y_{t+1} conditional on information up to time t.

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. Under adaptive learning (Sargent, 1993, Evans and Honkapohja, 2001), agents in the model are assumed to "act like statisticians or econometricians when doing the forecasting about the future state of the economy" (Evans and Honkapohja, 2001, p. 13). Specifically, agents form expectations based on some perceived law of motion (PLM)

for the process y_t, whose parameters are recursively estimated using information available to them. Their forecasts or learning algorithms can be expressed as weighted averages of past data, where the weights may vary over time to reflect information accrual as the sample increases, which is a key aspect of learning. For the purposes of this paper, we follow CM and restrict our attention to linear learning algorithms. Linear learning algorithms can be motivated using the so-called mean-plus-noise model as PLM, see equations (4) below. It is typical in the literature to assume that the PLM nests some rational expectations equilibrium, so that agents in the model are in some sense 'boundedly rational' (Sargent, 1993). This is not essential for our results so we assume that x_t is persistent but has short memory (see Assumption B below).

The simplest PLM for y_t is the mean-plus-noise model

$$y_t = \alpha + \epsilon_t, \tag{2}$$

where α is an unknown parameter, and ϵ_t is an identically and independently distributed (*i.i.d.*) shock.[1] Under this PLM, the conditional expectation of y_{t+1} given information up to time t is simply α, and because it is unknown to the agents, their forecast y^e_{t+1} is given by a recursive estimate of α. The classic learning algorithm is recursive least squares (RLS): $y^e_{t+1} = \frac{1}{t}\sum_{i=1}^{t} y_i$. This is a member of the class of weighted least squares algorithms that are defined as the solution to the minimization problem:

$$y^e_{t+1} = \underset{a}{\operatorname{argmin}} \sum_{j=0}^{t-1} w_{t,j} \left(y_{t-j} - a\right)^2, \quad \sum_{j=0}^{t-1} w_{t,j} = 1. \tag{3}$$

RLS corresponds to $w_{t,j} = t^{-1}$. Another member of this class, which is particularly popular in applied work, obtains when the weights decline exponentially, i.e., $w_{t,j} \propto (1-\bar{g})^j$ for some constant $\bar{g} \in (0,1)$.

An alternative characterization of learning in the literature is based on stochastic recursive algorithms (see Evans and Honkapohja, 2001, Chapter 6). Consider a slight generalization of the PLM (2) to allow for *perceived* shifts in the mean:

$$y_t = \alpha_t + \epsilon_t, \tag{4a}$$

$$\alpha_t = \alpha_{t-1} + v_t, \quad t \geq 1, \tag{4b}$$

where $\alpha_0 = \alpha$; ϵ_t and v_t are *i.i.d.* with mean zero and finite variances, and define the signal-to-noise ratio $\tau_t = \mathsf{Var}\left(v_t\right) / \mathsf{Var}\left(\epsilon_t\right)$. Under the PLM (4), y^e_{t+1} is given by a function of current and past values of y_t that estimates α_t. If the errors ϵ_t, v_t are Gaussian, the optimal estimate

[1] This PLM nests the rational expectations equilibrium that arises when $E_t\left(x_{t+j}\right)$ is constant for all t, j. Otherwise, it can be interpreted as a restricted perceptions equilibrium (RPE), see Sargent (1993).

of α_t, denoted by a_t, is given by the Gaussian Kalman Filter (see Durbin and Koopman, 2008):

$$a_t = a_{t-1} + g_t (y_t - a_{t-1}), \quad t \geq 1, \tag{5a}$$

$$g_t = \frac{g_{t-1} + \tau_t}{1 + g_{t-1} + \tau_t}, \quad t \geq 2, \quad g_1 = \frac{\sigma_0^2 + \tau_1}{1 + \sigma_0^2 + \tau_1} \tag{5b}$$

with a_0 and σ_0^2 measuring the mean and 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. When $g_t = \bar{g}$ for all t, the algorithm is called constant gain least squares (CGLS). RLS arises as a special case when $\tau_t = 0$ for all t and $\sigma_0^2 \to \infty$, so that $g_t = 1/t$. This is a member of a more general class of decreasing gain least squares (DGLS) algorithms where $g_t \sim \theta t^{-\nu}$, with $\theta > 0$ and $\nu \in (0, 1]$, as discussed Evans and Honkapohja (2001, Chapter 7) and e.g. Malmendier and Nagel (2016).[2]

The above learning algorithms can be all expressed as linear functions of past values of y_t with possibly time-varying coefficients:

$$y_{t+1}^e = \sum_{j=0}^{t-1} \kappa_{t,j} y_{t-j} + \varphi_t, \tag{6}$$

where the term φ_t represents the impact of the initial beliefs. We define the polynomial κ_t such that $\kappa_t(L) = \sum_{j=0}^{t-1} \kappa_{t,j} L^j$ where L is the lag operator. CGLS and DGLS algorithms can be expressed as stochastic recursive algorithms which only use the latest information (timed t) to update a_{t-1} into a_t. By contrast, the general formulation (6) allows for learning algorithms which require all the available observations for updating beliefs. One such example is studied below where κ_t is constant, i.e. $\kappa_t(L) = \kappa(L)$ with weights $\kappa_{t,j} = \kappa_j$ that decay hyperbolically in j.

To quantify how much agents discount past observations when forming expectations, we use the mean lag of κ_t, which is defined as

$$m(\kappa_t) = \frac{1}{\kappa_t(1)} \sum_{j=1}^{t-1} j \kappa_{t,j}. \tag{7}$$

The magnitude of $m(\kappa_t)$ relative to the sample size can be used to measure the 'length' of the learning window. We show below that this drives the memory of the process that is induced by learning dynamics. The following definition provides our measure of the length of the learning window.

[2] An alternative extension to the PLM (4b) allows for a non unit autoregressive parameter $\rho \in (0, 1)$, as in $\alpha_t = \rho \alpha_{t-1} + \nu_t$. When $\tau \neq 0$, the CGLS learning algorithm becomes $a_t = (1 - \rho\bar{g}) a_{t-1} + \bar{g} y_t$. When $\tau = 0$, the learning algorithm is DGLS with a gain that decays exponentially fast.

Definition LW (length of learning window) *Suppose there exist scalars* $m_\kappa > 0$ *and* $\delta_\kappa \geq 0$ *such that* $m(\kappa_t) \sim m_\kappa t^{\delta_\kappa}$, *as* $t \to \infty$. *Then,* δ_κ *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.*

In the paper, we make the following assumptions about the general linear learning algorithm (6):

Assumption A.

A.1. κ_t is nonstochastic;

A.2. $\{\kappa_{t,j}\}$ is absolutely summable with $\kappa_t(1) \leq 1$ for all t;

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 could be relaxed to allow $\{\kappa_{t,j}\}$ to be stochastic, provided that it is independent of $\{x_t\}$, in which case our results would be conditional on almost all realizations of $\{\kappa_{t,j}\}$. It precludes cases in which $\kappa_{t,j}$ depends on lags of y_t (or x_t), such as when the PLM is an autoregressive model, because in those cases the learning algorithm is nonlinear.[3]

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 but simplifies the exposition since $\delta_\kappa = 0$ implies here that $m(\kappa_t)$ is bounded.

We list the learning algorithms we study later in the paper in Table 1, where we also specify the length of the learning window for each algorithm. The first two algorithms are DGLS, and they were analyzed in CM. Both are long window algorithms. The next two algorithms are CGLS, discussed in Section 3.1. The last set of algorithms are weighted least squares algorithms with hyperbolically decaying weights, analyzed in Section 3.2.

Our working definition of long memory is the same as in CM, which applies both to stationary as well as non-stationary time series.[4]

Definition LM (long memory) *A second-order process* z_t *exhibits long memory of degree* $d > 0$ *if there exists* d such that

$$\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} z_t\right) \asymp T^d. \tag{8}$$

The process exhibits short memory if $d = 0$.

[3] Assumption A.1 also avoids the issue of generating fat tails through a random coefficient autoregressive model as in Benhabib and Dave (2014).

[4] 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 (8) above, see also Berenguer-Rico and Gonzalo (2014).

Learning Algorithm		$\kappa_{j,t}$	gain	δ_κ
DGLS	$\theta \geq 1$	$\theta \frac{\Gamma(t+1-\theta)}{\Gamma(t-j+1-\theta)} \frac{\Gamma(t-j)}{\Gamma(t+1)}$	$\min\left(\frac{\theta}{t}, 1\right)$	1
RLS	$\theta = 1$	$t^{-1} 1_{\{j<t\}}$	$\frac{1}{t}$	1
CGLS	$\bar{g} \in (0,1)$	$\bar{g}(1-\bar{g})^j$	$\bar{g}$	0
CGLS with small gain	$\bar{g}_T = c_g T^{-\lambda}$	$\bar{g}_T (1-\bar{g}_T)^j$	$\bar{g}_T$	λ
HWLS	$\lambda < 1$	$j^{\lambda-2}/\zeta(2-\lambda)$	-	$\max(0, \lambda)$
	$\lambda \in (1,2)$	$(\lambda-1) j^{\lambda-2} t^{1-\lambda}$	-	1

Table 1: Examples of Weighted Least-Squares Learning algorithms, with corresponding coefficients ($\kappa_{t,j}$), gains and learning window lengths (δ_κ). $\zeta(\cdot)$ denotes Riemann's Zeta function and $\Gamma(\cdot)$ is the Gamma function. DGLS: Decreasing Gain Least Squares; RLS: Recursive Least Squares; CGLS: Constant Gain Least Squares; HWLS: Hyperbolically Weighted Least Squares.

The above definition applies generally to any stochastic process that has finite second moments (which we assume in this paper). For a covariance stationary process, where the autocorrelation function (ACF) 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 (typically if they decay hyperbolically), or the spectrum has a pole at frequency zero. This gives rise to the following alternative definitions of d based on the ACF and spectral density that are equivalent to definition LM for covariance stationary processes, see Beran (1994) or Baillie (1996):

$$\begin{aligned} &\rho_z(k) \sim c_\rho k^{2d-1}, \quad &&\text{as } k \to \infty \\ &f_z(\omega) \sim c_f |\omega|^{-2d}, \quad &&\text{as } \omega \to 0, \end{aligned} \tag{9}$$

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 $\mathsf{I}(d)$, 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, $\mathsf{I}(d)$ processes admit a covariance stationary representation when $d \in (-1/2, 1/2)$, and are non-stationary if $d \geq 1/2$. Long memory arises when the degree of fractional integration is positive, $d > 0$. In the case of nonstationary

processes, the ACF definition of d in (9) does not apply,[5] so we use the ACF/spectrum of Δz, as in Heyde and Yang (1997):

$$\begin{aligned} &\rho_{\Delta z}(k) \sim c_\rho k^{2(d-1)-1}, \quad 1/2 < d < 1 \qquad \text{as } k \to \infty; \\ &f_{\Delta z}(\omega) \sim c_f |\omega|^{-2(d-1)}, \quad 1/2 < d < 1 \qquad \text{as } \omega \to 0. \end{aligned} \tag{10}$$

Finally, we make the following assumption about the forcing variable x_t.

Assumption B. There exists an *i.i.d.* process ϵ_t with $E[\epsilon_t^r] < \infty$ for $r > 4$, and such that $x_t = \sum_{j=0}^{\infty} \vartheta_j \epsilon_{t-j}$, with $\sum_{j=0}^{\infty} \vartheta_j \neq 0$ and $\sum_{j=0}^{\infty} j |\vartheta_j| < \infty$.

Assumption B ensures that x_t has short memory. We make this assumption in order to use results from Stock (1994) and Magdalinos and Phillips (2009). This assumption includes all covariance stationary processes that admit a finite-order invertible autoregressive moving average (ARMA) representation, and therefore have exponentially decaying autocovariances, but it also includes more persistent short memory processes whose autocovariances decay at slower-than-exponential rates with differentiable spectral density at the origin (Stock, 1994).

3 Theoretical results

First, we show that long memory *cannot* arise in this model with perpetual learning whenever the coefficient on expectations in the model (1) is strictly less than 1. The following proposition formalizes this statement for the case of learning algorithms with time-invariant coefficients, $\kappa_{t,j} = \kappa_j$ and $\varphi_t = 0$ in (6). This corresponds to a situation when learning has started a long time ago, so the effect of initial beliefs has disappeared.

Proposition 1 *Consider the model $y_t = \beta y_{t+1}^e + x_t$, where $y_{t+1}^e = \kappa(L) y_t$ where $\kappa(\cdot)$ satisfies Assumption A, and x_t satisfies Assumption B. Suppose that β is bounded below unity, i.e. $\beta \leq 1 - \eta$ for some $\eta > 0$. Then, as $T \to \infty$,*

$$\mathsf{sd}\left(T^{-1/2} \sum_{t=1}^{T} y_t\right) = O(1). \tag{11}$$

This result holds irrespective of the length of the learning window. To see why, the spectral density of y_t, $f_y(\omega)$ is bounded at the origin: it satisfies $f_y(\omega) = \left|1 - \beta\kappa\left(e^{i\omega}\right)\right|^{-2} f_x(\omega) \to |1 - \beta\kappa(1)|^{-2} f_x(0) < \infty$ as $\omega \to 0^+$ since $\kappa(1) \leq 1$. Then (11) follows from Beran (1994, Theorem 2.2).

Next, we turn to the empirically relevant cases when β is close to unity. We can model the persistence in such cases using local asymptotics that link the parameters to the observed

[5] 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.

sample, as was done by Chevillon *et al.* (2010) in a related paper. We extend their asymptotics by setting

$$\beta = 1 - c_\beta T^{-\nu}, \tag{12}$$

where c_β is a positive real number and $\nu \in [0, 1]$. This nests the case of fixed β in Proposition 1 when $\nu = 0$, and the specific asymptotic approximation in Chevillon *et al.* (2010) who used $\nu = 1/2$.[6] We will show that perpetual learning can lead to *apparent* long memory when $\nu > 0$, depending on the length of the learning window. Subsection 3.1 shows this for CGLS algorithms and subsection 3.2 shows it for hyperbolically weighted least squares algorithms.

3.1 Constant Gain Least Squares

For fixed gain, CGLS is clearly a short-window algorithm, but this is not an appropriate characterization when the gain parameter is small relative to the sample size, as is common in applied work, see e.g., Bullard and Eusepi (2005), Milani (2007), Eusepi and Preston (2011). To accommodate this case, we consider a local-to-zero asymptotic nesting where the gain parameter can go to zero with the sample size, i.e.,

$$\bar{g} = c_g T^{-\delta}, \tag{13}$$

where c_g is a positive real number and $\delta \in [0, 1]$. This nests the fixed gain case with $\delta = 0$, but can accommodate small-gain algorithms that mimic the behavior of long-window learning algorithms, which we study in the next subsection. We show in Appendix A.1 that the length of the learning window of the CGLS algorithm with $\bar{g}$ given by (13) is equal to δ. Allowing for a range of ν and δ allows us to provide an asymptotic approximation to the persistence of the data associated with different values of β and the length of the learning window that would not be possible if we kept them fixed at specific values as in Chevillon *et al.* (2010).

The CGLS algorithm on the mean-plus-noise PLM (4) makes y^e_{t+1} an exponentially weighted moving average of past y_j, $j \leq t$. Specifically, $y^e_{t+1} = a_t$, where

$$a_t = \left(\frac{1-\bar{g}}{1-\beta\bar{g}}\right)^t a_0 + \frac{\bar{g}}{1-\beta\bar{g}} \sum_{i=0}^{t-1} \left(\frac{1-\bar{g}}{1-\beta\bar{g}}\right)^i x_{t-i}. \tag{14}$$

To characterize the dynamics of y_t when $1-\beta$ and $\bar{g}$ may be close to zero, we use (12) and (13). Formally, this framework 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, $\bar{g}$ and y_t on T for notational simplicity. Assuming $(\nu, \delta) \in [0, 1]^2$ may lead $(1-\bar{g})/(1-\beta\bar{g})$ to lie outside an $O\left(T^{-1}\right)$ neighborhood of unity typically considered in the time-series literature. Our results hence relate to the work by

[6] West (2012) also used this local asymptotic parameterization with $\nu = 1/2$ in a model with rational expectations.

Giraitis and Phillips (2006), Phillips and Magdalinos (2007) and Andrews and Guggenberger (2007).

The following theorem gives the implications for the memory of y_t.

Theorem 2 *Consider the model $y_t = \beta y_{t+1}^e + x_t$, where $y_{t+1}^e = a_t$ given by (14), $a_0 = O_p(1)$ and x_t satisfies Assumption B. Suppose that $\beta = 1 - c_\beta T^{-\nu}$ and $\bar{g} = c_g T^{-\delta}$, where $\nu, \delta \in [0,1]^2$ and c_β, c_g are positive constants. Then, as $T \to \infty$,*

$$\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right) \asymp T^{\min(\nu, 1-\delta)}. \tag{15}$$

Theorem 2 shows that CGLS learning with a large β generates apparent long memory. More specifically, 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 $\nu = 0$, *i.e.*, β is 'far' from unity, the process exhibits short memory, irrespective of the length of the learning window. For $\nu > 0$, the memory of the process depends on whether $\nu \leq 1 - \delta$ or $\nu > 1 - \delta$, *i.e.*, on how close β is to unity relative to the length of the learning window. When β is sufficiently close to unity, the memory of the process is determined entirely by the length of the learning window, δ, and is nonincreasing in δ. Persistence is, in fact, strongest when the gain is far from zero, $\delta = 0$, *i.e.*, when the learning window is short. This may appear counterintuitive at first, but it is entirely analogous to what happens in fractionally integrated processes. To gain some intuition, consider the fractional white noise process $(1-L)^d z_t = \varepsilon_t$, where $d \in (-1/2, 1/2)$, $d \neq 0$, and ε_t is white noise. The memory of this process, d, is directly related to the rate of decay of the impulse response function, *i.e.*, the rate of decay of the coefficients of the moving average representation, which is $d - 1$.[7] The rate of decay of the autoregressive coefficients is $-d - 1$, so it is *inversely* related to d. Therefore, given a unit root in the autoregressive polynomial, a more persistent process is associated with a faster decay of the autoregressive coefficients. In the learning model, this corresponds to a higher discounting of past observations in the learning algorithm, *i.e.*, a shorter learning window.

CGLS learning with a small gain parameter induces behavior that is in some sense close to a rational expectations equilibrium, and it is referred to as 'near-rational expectations' in the literature, see Milani (2007). The smallest gain arises when $\delta = 1$ in Theorem 2, which leads to short memory. This is exactly what happens under rational expectations, see Proposition 1 in CM. So, similarly to rational expectations, learning that is akin to near-rational expectations cannot generate long memory.

Note that CGLS with very small gain is very different from RLS, *i.e.*, the latter is not the limit of the former as the gain parameter goes to zero. Heuristically, near-rational expectations corresponds to the 'limiting' law of motion when RLS learning has converged, and

[7] See, e.g., Baillie (1996, Table 2).

therefore, it misses all the transitional dynamics of RLS, which matter – this is exactly the intuition behind Theorem 2 in CM.

Sargent (1999, ch. 6) studies the model that we analyze in this paper with CGLS learning. Sargent observes that the PLM (4) has a unit root, while the actual law of motion (ALM) is a stationary ARMA process, and focuses on the misspecification of agents' beliefs. He estimates near self-confirming parameter values using frequency domain MLE and shows that the actual and perceived spectral densities differ at very low frequencies. In Sargent's example, the gain parameter is small, but the coefficient on expectations is far from unity ($\beta = 0.5$). Our results show that the ALM exhibits apparent long memory when β is close to one, which implies that the difference between the ALM and the PLM at low frequencies will be even smaller than in Sargent's example.

3.2 Hyperbolically Weighted Least Squares

In this subsection, we cover long-window learning algorithms (6) that satisfy Assumption A and have constant coefficients $\kappa_{t,j} = \kappa_j$. If we set initial beliefs appropriately, CGLS is such an algorithm, but without making the gain parameter local to zero, the weights κ_j decay exponentially and the length of the learning window is short. We now consider situations when weights of the learning algorithm decay hyperbolically in j, so that we can cover long-window algorithms without treating the gain parameter as local to zero. Such algorithms can be motivated as hyperbolically discounted, or weighted, least squares (HWLS). In some sense, they bridge the gap between RLS (no discounting) and CGLS (exponential discounting). 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 $\delta_\kappa \in (0, 2)$, the learning window is long, with length δ_κ.

One example of $\kappa(L)$ that satisfies the above assumptions is the operator $L_g = 1 - (1 - L)^g$, $g \in (0, 1)$, such that $\kappa_j \sim c_\kappa j^{-g-1}$, and $\delta_\kappa = 1 - g$, see Granger (1986) and Johansen (2008). This specific algorithm constitutes the optimal method for forming expectations about y_{t+1} is agents believe the process to be fractionally integrated of order g. In fact, we show in the next subsection that such beliefs can be self-confirming.

As in the case of CGLS, we use (12) and suppress the triangular array notation for y_t. Unlike CGLS, the weights of the learning algorithm here do not depend on T.

The following result gives the memory properties of the process y_t according to Definition LM.

Theorem 3 *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 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* $T \to \infty$,

$$\mathsf{sd}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right) \asymp T^{\min(\nu, 1-\delta_\kappa)}.$$

This result is entirely analogous to Theorem 2, where $\delta_\kappa = \delta$. 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 common definitions of long memory. These definitions rely either on the hyperbolic behavior of the spectral density in a neighborhood of the origin or on hyperbolic rates of decay of the autocorrelations.

Alternative characterizations of long memory also hold in this context as the following theorem shows.

Theorem 4 *Under the assumptions of Theorem 3, 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* $j = 1, ...n$, *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* $T, k \to \infty$,

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

The theorem shows that the degree of memory measured in Theorem 3 through Definition LM coincides with common alternative definitions.

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.

3.3 Learning to believe in long memory

Having established the apparent long memory implications of perpetual learning dynamics, we now turn to the properties of the Geweke and Porter-Hudak (1983), GPH, estimator of the long memory parameter d when agents learn about y_t. We rely on the high level assumption that there exists an estimator of the spectral density 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 5 *Under the model and assumptions of theorem 3 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, $j = 1, ..., n$: *as* $T \to \infty$,
if $\delta_\kappa \in (1/2, 1)$, $\widehat{f}_{y,T}(\omega_j) / f_y(\omega_j) \xrightarrow{p} 1$, *or if* $\delta_\kappa \in (0, 1/2)$, $\widehat{f}_{\Delta y,T}(\omega_j) / f_{\Delta y}(\omega_j) \xrightarrow{p} 1$.
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 Hommes and Sorger (1998) and Cho and Sargent (2008). If agents possess an ex ante belief that the process y_t exhibits long memory of degree d in the form of fractional integration they optimally learn using the hyperbolically weighted filter $\kappa(L) = 1 - (1 - L)^d$ with window $\delta_\kappa = 1 - d$. If β is sufficiently close to unity, then the data which is generated through agents' learning exhibits long memory of degree d. Agents who estimate the degree of long memory using the GPH estimator find a value $\widehat{d}$ which converges asymptotically to their ex ante belief, hence confirming it. This result relates to Hommes and Sorger (1998) who consider consistent expectations equilibria in deterministic sequences and where agents estimate sample autocorrelations. Here we show that such a mechanism holds in a stochastic setting where agents estimate the degree of persistence of the process.

Further discussion of self-confirming equilibria. We can think of self-confirming equilibria as "fixed points of mappings from beliefs to population limits of statistical models" (Sargent, 1999, p. 69). In the present context, the statistical object of interest is the memory parameter, d, of the time series. We consider the situation in which agents believe that the data has long memory of degree d, and accordingly, choose a forecasting model with weights that decay hyperbolically at rate $-d-1$ (e.g., ARFIMA(p, d, q) for any p, q). Let $\bar{d}(d)$ be the population limit of the memory parameter of the resulting ALM. If $\bar{d}(d) = d$, then agents' beliefs are self-confirming. Theorems 3 and 4 show that any HWLS algorithm with learning window length $\delta_\kappa \in (0, 1)$ can be a supported as a self-confirming equilibrium if β is sufficiently close to unity (in a sense that is made precise in terms of the rates $\nu > 1 - \delta_\kappa$).

There has been a lot of work in the learning literature discussing self-confirming equilibria in both the I(0) and I(1) extreme cases. The results here can be used to explore self-confirming equilibria in the middle ground of fractional integration. One way to do this would be to

discuss stability, for example, whether a self-confirming degree of persistence can be learned in real-time, and how agents would learn d in practice. Theorem 5 represents a first step in that direction, by establishing the limiting behavior of econometric estimates of d based on real time data. It suggests that a self-confirming degree of memory could be learnable, but it does not address the issue of how agents would learn it in practice (for example, in Theorem 5, agents beliefs about d are not updated, so Theorem 5 says that if agents start with a given set of beliefs, then the ALM will eventually confirm them, but it doesn't say what will happen if they update their beliefs in real time). This is an interesting topic for further research.

4 Simulations

This section presents simulation evidence in support of the analytical results given above. We generate samples of $\{y_t\}$ from (1) under the learning algorithms listed in Table 1 (we also report DGLS for completeness). The exogenous variable x_t is assumed to be *i.i.d.* normal with mean zero, and its variance is normalized to 1 without loss of generality. We use a small sample of size $T = 100$ and various values of the parameters β and $\bar{g}$ (results over a longer sample size $T = 1000$ are reported in the online supplement). We study the behavior of the variance of partial sums, the spectral density, and the popular Geweke and Porter-Hudak (1983) (henceforth GPH) and the Robinson (1995) maximum local Whittle likelihood estimators of the fractional differencing parameter d.[8] We also report the power of tests of the null hypotheses $d = 0$ and $d = 1$. The number of Monte Carlo replications is 10,000.

Figure 1 reports the Monte Carlo average log sample periodogram against the log frequency ($\log \omega$) under RLS and CGLS learning. This constitutes a standard visual evaluation of the presence of long range dependence if the log periodogram is linearly decreasing in log ω. 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, as is shown in CM. Under CGLS, we observe also that apparent long memory arises when β gets closer to unity but that it depends on the value of $\bar{g}$. Table 3 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. The behavior of $E(\widehat{d})$ as well as $\Pr(\text{Reject } d = 0)$ and $\Pr(\text{Reject } d = 1)$ in terms of β and $\bar{g}$ accords with Theorem 2. Specifically, $E(\widehat{d})$ is increasing in β given $\bar{g}$, and weakly increasing in $\bar{g}$ given β. Since T is fixed, a higher $\bar{g}$ corresponds to a shorter learning window, so the memory of the process is decreasing in the length of the learning window, in accordance with Theorem 2.

[8] We use $n = \left\lfloor T^{1/2} \right\rfloor$ Fourier ordinates, where $\lfloor x \rfloor$ denote the integer part of x.

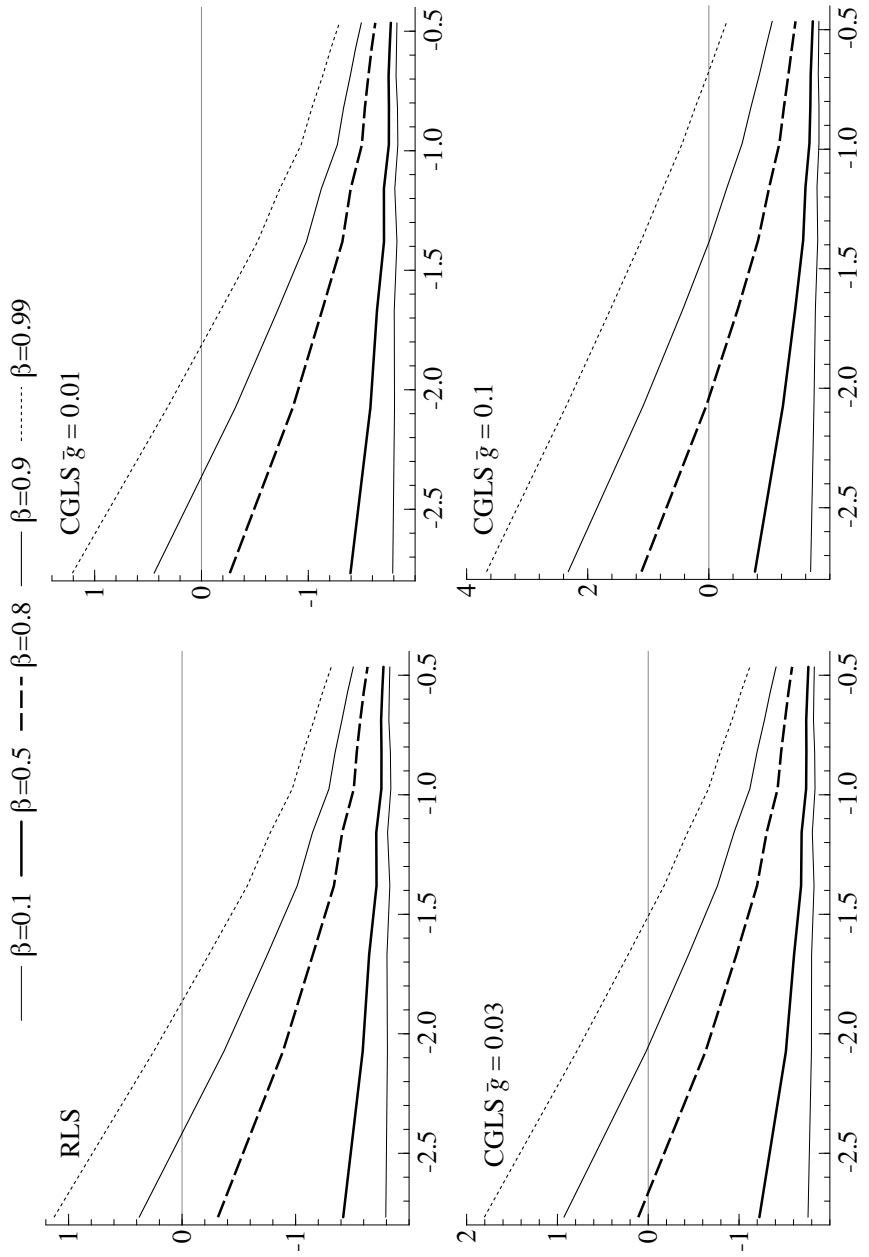

Figure 1: Monte Carlo averages over $10,000$ replications of the log periodogram against the log of the first $\sqrt{T}$ Fourier frequencies with $T = 1,000$ observations. The model is $y_t = \beta y_{t+1}^e + x_t$, $x_t \overset{i.i.d.}{\sim} \mathsf{N}(0,1)$, and y_{t+1}^e is determined by RLS (top left panel) or CGLS (all other panels) learning.

Figures 2 and 3 report the densities of the GPH and local Whittle likelihood estimators $\widehat{d}$ of the degree of fractional integration of y_t. The local Whittle estimator is obtained by constrained maximization over the range $d \in (-1, 2)$. The model is the 'mean plus noise' model of the paper and the simulation settings are the same as in Figure 1 and Table 3. Unreported simulations 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 learning with local parameters.

5 Application to Present Value Models

We now consider the implications of learning in the present value models of Campbell and Shiller (1987) for stock prices and Engel and West (2005) for exchange rates. Observed long memory in the dividend-price ratio and the forward premia have been used to explain the well-known empirical puzzles of excess return predictability by Schotman et al. (2008) and the forward premium anomaly by Baillie and Bollerslev (2000) and Maynard and Phillips (2001). Here we study whether present value models with learning can explain the long memory observed in the data.

There are some related papers that report results complementary to ours. Benhabib and Dave (2014) studied models for asset prices and show that some forms of learning may generate a power law for the distribution of the log dividend-price ratio. Branch and Evans (2010), and Chakraborty and Evans (2008) studied the potential of adaptive learning to explain the empirical puzzles. The former focus on explaining regime-switching in returns and their volatility, rather than low frequency properties of the dividend-price ratio, and the latter assume that fundamentals are strongly persistent (near random walks).

There are several papers in the literature that rely on the proximity of the coefficient on expectations to unity in order to explain the above empirical puzzles. Engel and West (2004, 2005) and Engel et al. (2008) do so in the context of rational expectations whereas Chakraborty and Evans (2008) do so in the context of adaptive learning. All of these papers rely on the fundamentals being persistent (near or exact unit roots) and the persistence being of exogenous origin. In contrast, the present paper focuses on the situation where the fundamentals do not exhibit any exogenous persistence, so the observed persistence in the endogenous variable arises solely from adaptive learning.

5.1 Stock prices

Let P_t, D_t and r_t denote the price, dividend and excess return, respectively, of an index of stocks. Under the rational expectations asset pricing model of Campbell and Shiller (1988),

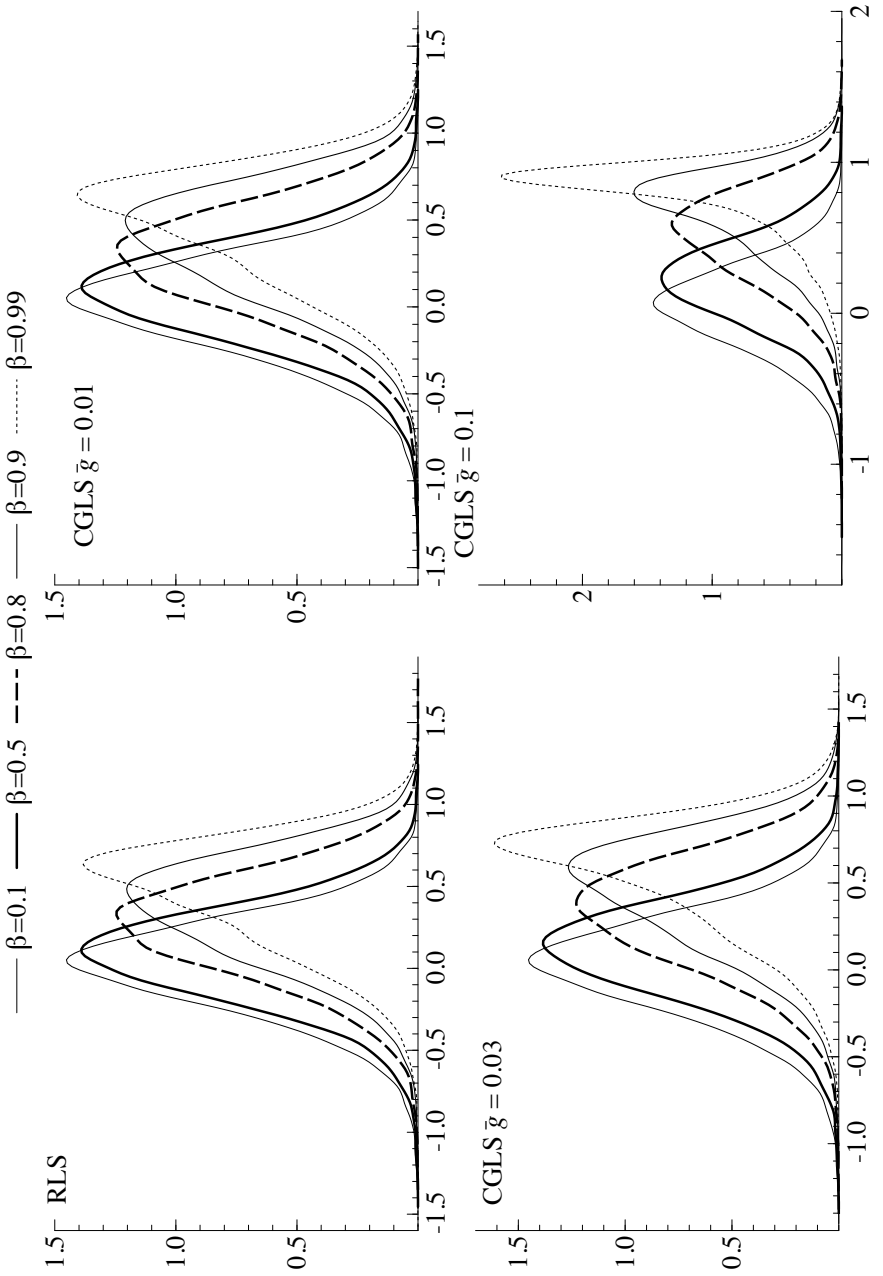

Figure 2: Density of the GPH log periodogram estimator of the fractional integration parameter d using $n = \sqrt{T}$ Fourier frequencies over samples of $T = 100$ observations. The model is the 'mean plus noise' perceived law of motion presented in Section 2. g is 0 under RLS learning and $\bar{g}$ otherwise. The number of Monte Carlo replications is 10,000.

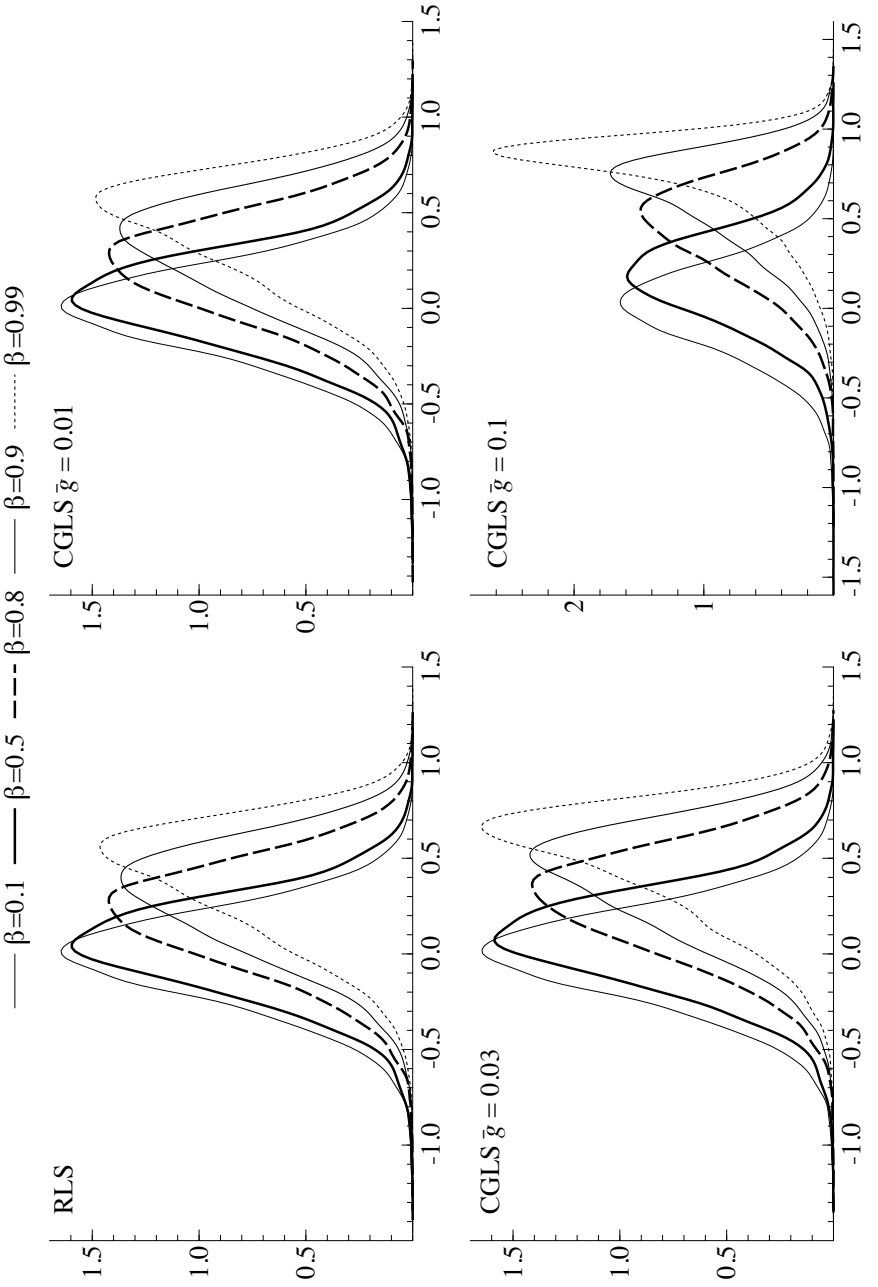

Figure 3: Density of the Whittle likelihood estimator of the fractional integration parameter d using $n = \sqrt{T}$ Fourier frequencies over samples of $T = 100$ observations. The model is the 'mean plus noise' perceived law of motion presented in Section 2. g is 0 under RLS learning and $\bar{g}$ otherwise. The number of Monte Carlo replications is 10,000.

the log dividend-price ratio is given by

$$\log \frac{D_t}{P_t} = c + E_t \sum_{j=0}^{\infty} \beta^j \left(-\Delta \log D_{t+j+1} + r_{t+j+1}\right), \tag{16}$$

where c, β are log-linearization parameters, see also Campbell, Lo and McKinlay (1996, equation 7.1.24). Equation (16) obtains as the bubble-free solution of the following first-order difference equation

$$\log \frac{D_t}{P_t} = (1-\beta)\, c + \beta E_t \left(\log \frac{D_{t+1}}{P_{t+1}}\right) + E_t \left(r_{t+1} - \Delta \log D_{t+1}\right). \tag{17}$$

The above equation can be written in the form (1) with $y_t = \log \frac{D_t}{P_t}$ and $x_t = (1-\beta)\, c + E_t \left(r_{t+1} - \Delta \log D_{t+1}\right)$. We have data on y_t, but we do not observe the driving process x_t, because it depends on *expected* returns and dividend growth which are unobserved. Proposition 1 in CM shows that if x_t exhibits short memory, then y_t should also exhibit short memory.

Figure 4 plots measures of $\log\left(D_t/P_t\right)$, r_t and $\Delta \log D_t$ using annual data on the Standard and Poor's (S&P) stock index over the period 1871-2011 available from Robert Shiller's website. An apparently puzzling feature of the data is that the log dividend-price ratio exhibits very strong persistence, while dividend growth and excess returns show hardly any signs of persistence. This is demonstrated using two of the most recent estimators of the degree of memory which are both efficient and consistent under weak assumptions (Shimotsu and Phillips, 2005, Shimotsu 2010, and Abadir, Distaso and Giraitis, 2007), as reported in Panel A of Table 4. Both estimators show that y_t exhibits long memory with memory parameter 0.79 and 0.85, and significantly different from zero. In contrast, $\Delta \log D_t$ and r_t exhibit short memory, since the estimates of their memory parameters are zero, and this is consistent with our Assumption B that the fundamental process x_t exhibits short memory.[9]

We now turn to the question of whether it is possible to explain the observed low frequency variation in $\log\left(D_t/P_t\right)$ endogenously using learning, that is, when the exogenous process x_t exhibits short memory. In our empirical analysis, we calibrate β to 0.96, based on Campbell, Lo and McKinlay (1996, chapter 7, p. 261). For any given learning algorithm, characterized by some parameter ϑ, say, we compute the expectation under learning, denoted $y^e_{t+1}(\vartheta)$, and $x_t(\vartheta) = y_t - \beta y^e_{t+1}(\vartheta)$. We then test the null hypothesis that the memory parameter, d, of $x_t(\vartheta)$ is zero against a one-sided alternative that it is positive. We use one-sided t-tests based on the Shimotsu and Phillips (2005) and Abadir *et al.* (2007) estimators, as in Table 4. If

[9] It is theoretically possible that x_t contains a long-memory component that is so small as not to be statistically detectable. An extension of an argument in Campbell, Lo and McKinlay (1996, sec. 7.1.4) could be used to show that *realized* returns and dividend growth can *appear* to exhibit short memory even though *expected* returns and/or dividend growth may have a degree of long memory that is sufficient to explain the persistence in the log dividend-price ratio. Thus, one should not use the observed time series properties in Figure 4 and Table 4 as prima facie evidence against rational expectcations.

β	Mean of $\widehat{d}$ GPH	Whittle	Pr(Reject $d=0$) GPH	Whittle	Pr(Reject $d=1$) GPH	Whittle
0.10	0.007	-0.023	0.122	0.122	0.822	0.937
0.50	0.070	0.032	0.180	0.184	0.719	0.876
0.80	0.235	0.180	0.399	0.414	0.563	0.742
0.90	0.337	0.274	0.537	0.556	0.495	0.667
0.99	0.442	0.375	0.669	0.681	0.430	0.599

Table 2: The table records estimates and tests on the long memory d for $y_t = \beta y_{t+1}^e + x_t$, under RLS learning. The data is generated as $x_t \overset{i.i.d.}{\sim} \mathsf{N}(0,1)$, $T = 100$ and the number of Monte Carlo replications is 10,000. GPH is the Geweke & Porter-Hudak (1983) estimator and Whittle is the Robinson (1995) maximum local Whittle likelihood estimator. $\Pr(\text{Reject } d=0)$ and $\Pr(\text{Reject } d=1)$ are the empirical rejection frequencies of one-sided 5% level tests of $\mathsf{H}_0 : d = 0$ against $\mathsf{H}_1 : d > 0$, and $\mathsf{H}_0 : d = 1$ against $\mathsf{H}_1 : d < 1$, resp.

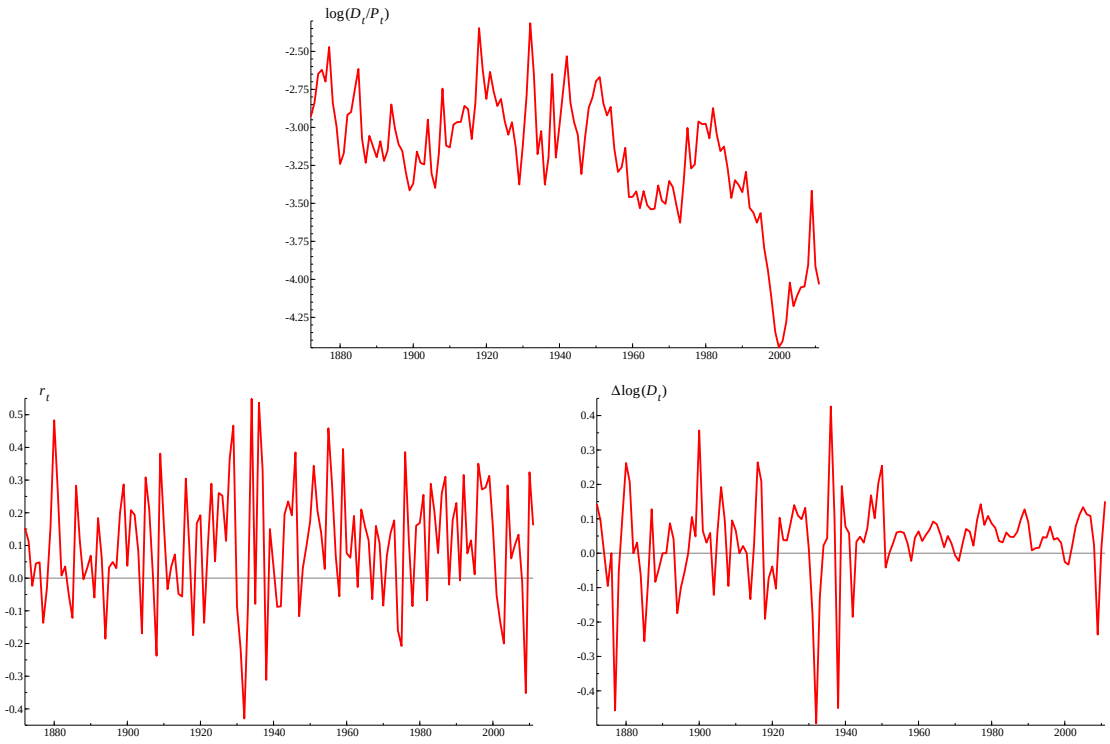

Figure 4: Log dividend-price ratio, returns and dividend growth for S&P annual index data.

		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	0.007	-0.022	0.122	0.123	0.823	0.937
	0.50	0.074	0.036	0.186	0.190	0.719	0.876
	0.80	0.243	0.189	0.411	0.427	0.565	0.743
	0.90	0.347	0.285	0.552	0.568	0.498	0.668
	0.99	0.454	0.387	0.681	0.696	0.433	0.601
0.03	0.10	0.013	-0.017	0.127	0.125	0.823	0.937
	0.50	0.104	0.066	0.217	0.230	0.716	0.876
	0.80	0.306	0.249	0.495	0.513	0.568	0.750
	0.90	0.425	0.360	0.639	0.656	0.507	0.680
	0.99	0.541	0.474	0.762	0.777	0.443	0.614
0.10	0.10	0.029	-0.001	0.139	0.141	0.821	0.935
	0.50	0.195	0.157	0.329	0.360	0.707	0.866
	0.80	0.487	0.428	0.699	0.738	0.551	0.734
	0.90	0.645	0.581	0.836	0.859	0.505	0.688
	0.99	0.781	0.725	0.918	0.934	0.478	0.673

Table 3: The table records estimates and tests on the long memory d for $y_t = \beta y^e_{t+1} + x_t$, under CGLS learning with gain parameter $\bar{g}$. The data is generated as $x_t \overset{i.i.d.}{\sim} \mathsf{N}(0,1)$, $T = 100$ and the number of Monte Carlo replications is 10000. GPH is the Geweke & Porter-Hudak (1983) estimator and Whittle is the Robinson (1995) maximum local Whittle likelihood estimator. $\Pr(\text{Reject } d=0)$ and $\Pr(\text{Reject } d=1)$ are the empirical rejection frequencies of one-sided 5% level tests of $\mathsf{H}_0 : d = 0$ against $\mathsf{H}_1 : d > 0$, and $\mathsf{H}_0 : d = 1$ against $\mathsf{H}_1 : d < 1$, resp.

there is a value of ϑ for which the test does not reject the null hypothesis, we can conclude that there is a learning algorithm of the type indexed by ϑ that can explain the low frequency variation in y_t. This strategy provides a formal test of the fit of the model, and the least rejected value of ϑ constitutes a Hodges and Lehmann (1963) estimate.

We consider the two classes of learning algorithms studied earlier: CGLS, with $\vartheta = \bar{g} \in (0,1)$; and DGLS, with $\vartheta = \theta \in [1,5]$. Theorem 2 implies that, when β is close to one, the memory of y_t is increasing in $\bar{g}$, so we report the minimum value of $\bar{g}$ for which the null hypothesis is not rejected, *i.e.*, the minimum value of $\bar{g}$ that is consistent with the memory of y_t under CGLS learning when x_t has short memory. The results for $\log(D_t/P_t)$ are given in the first column of Table 5. Both tests yield similar values of $\bar{g} = 0.23$ and 0.24.[10] Next, we turn to DGLS algorithms covered in CM. We find that there is no value of θ for which the null hypothesis is accepted, so we conclude that DGLS learning dynamics (including RLS), under the PLM considered, do not match the low frequency variation in the data.

Finally, we consider learning algorithm with hyperbolic weights such that $y^e_{t+1} = \kappa(L)\, y_t$ with $\kappa(L) = 1-(1-L)^g$ for $g \in (0,1)$ so $\delta_\kappa = 1-g$. The first column of Table 6 reports the minimum parameter g for which the null hypothesis is not rejected, *i.e.*, the minimum value of g that is consistent with the memory of y_t under hyperbolically discounted least squares learning when x_t has short memory. The values of g thus obtained (0.61 and 0.64 depending on the estimator) corresponds to a relatively low value of δ_κ.

5.2 Exchange rates

The forward premium anomaly constitutes another puzzling empirical feature that is related to present value models and has been explained via long memory, see Maynard and Phillips (2001). The puzzle originates from the Uncovered Interest Parity (UIP) equation:

$$E_t[s_{t+1} - s_t] = f_t - s_t = i_t - i^*_t \tag{18}$$

where s_t is the log spot exchange rate, f_t 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 and the second equality follows from the covered interest parity. The UIP under the efficient markets hypothesis has been tested since Fama (1984) as the null $\mathsf{H}_0 : (c,\gamma) = (0,1)$ in the regression

$$\Delta s_t = c + \gamma(f_{t-1} - s_{t-1}) + \epsilon_t. \tag{19}$$

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

Baillie and Bollerslev (2000) and Maynard and Phillips (2001) suggest econometric explanations of this puzzle that rely on strong persistence of the forward premium. Baillie and

[10] Benhabib and Dave (2014) report estimates of the gain parameter of that order of magnitude. They identify the gain though the implied tail distribution of y_t.

Panel A: Stock prices and dividends

Estimator	$\log(D_t/P_t)$	r	$\Delta \log(D_t)$
2ELW	0.85	0.13	0.11
FELW	0.79	0.13	0.05
s.e.	0.15	0.15	0.15

Panel B: Forward premia

Estimator	Canada	France	Germany	Italy	Japan	UK
$\hat{d}_{2ELW}$	0.52	0.43	0.80	0.75	0.63	0.65
$\hat{d}_{FELW}$	0.50	0.50	0.80	0.68	0.63	0.50
s.e.	0.14	0.14	0.14	0.15	0.15	0.14
Sample size	151	151	151	138	137	151

Table 4: Estimates of the degree of long memory. 2ELW is the Two-Step Exact Whittle Likelihood Estimator of Shimotsu and Phillips (2005) and Shimotsu (2010), FELW is the Nonstationary-Extended local Whittle estimator of Abadir et al. (2007). Standard errors are the same for both estimators. Panel A corresponds to annual S&P data since 1871. Panel B corresponds to quarterly Eurodollar interest differentials for each of the indicated currencies from the mid-1970s.

	$\log(D_t/P_t)$	Canada	France	Germany	Italy	Japan	UK
2ELW	0.26	0.11	0.04	0.20	0.15	0.12	0.08
FELW	0.26	0.12	0.04	0.21	0.15	0.12	0.08

Table 5: The table reports the minimum value of the gain parameter such that a t-test of $\mathsf{H}_0 : d = 0$ versus $\mathsf{H}_1 : d > 0$ is not rejected for $x_t = y_t - \beta y^e_{t+1}$ at a 5% asymptotic nominal level of significance. For details of estimators and data, see Table 4.

Bollerslev (2000) provide "evidence that this so-called anomaly may be viewed mainly as a statistical phenomenon that occurs because of the very persistent autocorrelation in the forward premium." Their explanation is based on persistent volatility. Maynard and Phillips (2001) show that if the forward premium $i_t - i_t^*$ is fractionally integrated and Δs_t is a short memory process that satisfies our Assumption B, then OLS estimates of γ in (19) converge to zero and have considerable probability of being negative in finite samples. They provide evidence of long memory in forward premia for several countries relative to the US dollar. We look at the data on three-month Eurodollar interest differentials for six countries, Canada, France, Germany, Italy, Japan and the UK, over the period ranging from the mid-1970s to 2012 (starting points vary by country). The dataset is the one used by Engel and West (2005), updated from Thomson Datastream.[11] Figure 5 plots the time series, and Panel B of Table 4 provides estimates of their memory parameters. We see that all series exhibit strong persistence with estimates of d greater than 0.4, corroborating the results in Maynard and Phillips (2001).

A possible explanation for the strong persistence in the forward premium is the presence of an exogenous time-varying risk premium, see Engel (1996). Under this explanation, the UIP equation becomes

$$E_t \left[s_{t+1} - s_t \right] = i_t - i_t^* + \rho_t, \tag{20}$$

where ρ_t is an unobserved process that represents a time-varying risk premium. In order to match the long memory of the forward premia under rational expectations, the exogenous risk premium ρ_t must exhibit long memory, too, since Δs_t appears close to *i.i.d.*, see Engel and West (2005).

We investigate whether learning dynamics can generate enough persistence to match the low frequency variation in the forward premia, without assuming that it arises exogenously through the risk premium. We consider the two exchange rate models studied in Engel and West (2005), a money-income model with an exogenous real exchange rate, and a Taylor rule model where the foreign country has an explicit exchange rate target. We show that each of these models implies a forward-looking equation for the forward premium $y_t = i_t - i_t^*$ of the form (1), with a different driving process x_t, and a different interpretation of the coefficient β for each model (derivations are given in Section I of the Appendix). Specifically, letting z_t denote a vector of 'fundamentals' that includes money, income, price and inflation differentials, the real exchange rate, and a nominal exchange rate target, it can be shown that y_t follows (1) with $x_t = (1-\beta)\left(b' E_t \Delta z_{t+1} - \rho_t\right)$, where b is a vector of coefficients that depends on the model. In the money income model, β is a function of the interest semi-elasticity of money demand, while in the Taylor rule model, β is inversely related to the

[11] Available from http://www.ssc.wisc.edu/~cengel/Data/Fundamentals/data.htm and Datastream under mnemonics S20520, S20544, S20544, S98803, S20963, S20508 and for the US: S20514.

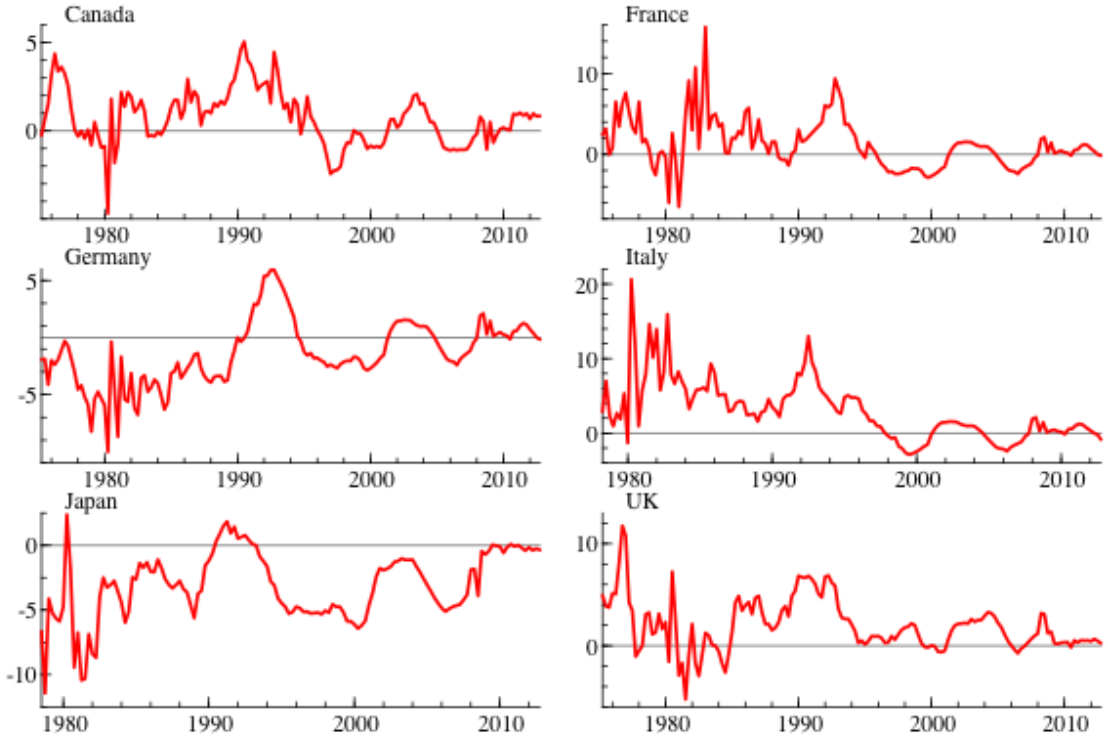

Figure 5: Forward premia with respect to the US dollar for six countries.

degree of the intervention of foreign monetary authorities to target the exchange rate. Using past empirical studies, Engel and West (2005) calibrate β within the range $0.97-0.98$ for the money income model and $0.975-0.988$ for the Taylor rule model. For the empirical analysis here we choose the value $\beta = 0.98$, which covers both models.

We perform the same analysis as in the previous subsection, to identify any learning algorithms that can explain the persistence in y_t when x_t is short memory. The results are entirely analogous to the case of the dividend-price ratio. Specifically, we find no DGLS learning algorithm that can explain the long memory in the forward premia when the fundamentals have short memory, but we do find CGLS learning algorithms that can. The minimum gain parameters needed for each country are reported in columns 2-7 of Table 5. The smallest gain parameter corresponds to France (0.04), and the largest to Germany (0.21). These gains are somewhat higher than the values typically used in the applied learning literature, see, *e.g.*, Chakraborty and Evans (2008) for this application.

Turning to HWLS, Table 6 reports as before the minimum parameter g for which the null hypothesis is not rejected. As with CGLS The minimum g that satisfy this property tends to be smaller in this example than in the Campbell-Shiller setting (to the exception of Germany and Italy). All in all, our conclusions are analogous to the case of the dividend-price ratio.

	$\log(D_t/P_t)$	Canada	France	Germany	Italy	Japan	UK
2ELW	0.61	0.36	0.22	0.62	0.63	0.55	0.15
FELW	0.64	0.38	0.20	0.63	0.61	0.51	0.07

Table 6: The table reports the minimum value of the parameter g such that a t-test of $\mathsf{H}_0 : d = 0$ versus $\mathsf{H}_1 : d > 0$ is not rejected for $x_t = y_t - \beta y^e_{t+1}$, where $y^e_{t+1} = (1-(1-L)^g)\, y_t$, at a 5% asymptotic nominal level of significance. For details of estimators and data, see Table 4.

There are large and growing literatures on the return predictability puzzle and the forward premium anomaly, see Cochrane (2011), Campbell (2014) and Engel (2014, 2016) for recent surveys. Our results contribute to the part of the literature that attributes the puzzles to "statistical bias" (Engel, 2014, Phillips, 2015), because they provide an explanation for the apparent long memory in the regressors (expected log dividend-price ratio or interest differentials) that has in turn been used to explain why econometric evidence may be unreliable (Baillie and Bollerslev, 2000, Maynard and Phillips, 2001). Most other papers in these literatures work under the assumption of rational expectations and propose alternative explanations of the puzzles. These are based on time-varying risk premia, liquidity premia, peso problems and stochastic discount factors. No consensus seems to have emerged so far, and many of the proposed explanations have led to new puzzles, such as the excess volatility of the exchange rate, see Engel (2016) or new empirical evidence, such as the role of tail risk,

the so-called *fear* factors (Bollerslev, Todorov and Xu, 2015).

6 Conclusion

We studied the implications of learning in models where endogenous variables depend on agents' expectations and complement the results of CM who studied the persistence induced under Decreasing Gain Least-Squares Learning. In a prototypical representative-agent forward-looking model with linear learning algorithms, we find that learning can generate strong persistence under perpetual learning. The degree of persistence induced by learning depends negatively on the weight agents place on past observations when they update their beliefs, and positively on the magnitude of the feedback from expectations to the endogenous variable. In algorithms with shorter windows, long memory provides an approximation to the low-frequency variation of the endogenous variable. We also show that agents' beliefs in long memory can be self confirming. Finally, the apparent long memory induced by learning can shed some light on well-known empirical puzzles in present value models.

A Appendix

A.1 Proof of $\delta_\kappa = \delta$ under CGLS

Under CGLS learning the algorithm is

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

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

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

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

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

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

$$m(\kappa_T) \asymp T^\delta, \tag{21}$$

i.e., $\delta_\kappa = \delta$.

A.2 Preliminary Lemmas

We provide here some lemmas which will be useful in the subsequent proofs. The proofs are in the Supplementary Appendix.

Lemma 6 *Let f a spectral density with f, f' and f'' bounded, $f > 0$ in a neighborhood of the origin and $f'(0) = 0$. Let $|\delta| \in (0,1)$ and $\omega_k = 2\pi k/T$, $k = o(T)$. Then,*

$$T^{\delta-1} \sum_{j=1}^{T} j^{-\delta} f(\omega_j) \cos(j\omega_k) \asymp k^{\delta-1}. \tag{22}$$

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 $\delta_\kappa \in (0,1)$. Assume $\kappa(1) = 1$. Then, there exist $c_\kappa^* \neq 0$ and $c_\kappa^{**} > 0$ such that*

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

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 $\delta_\kappa \in (0,1)$. We assume that $\beta \leq \kappa(1) - \eta$ for some $\eta > 0$ and let f_y denote the spectral density of y_t. Then $f_y(0) < \infty$ and there exists $c_f > 0$ such that*

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

A.3 Proof of Theorem 2

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, $1 - \frac{(1-\beta)\bar{g}}{1-\beta\bar{g}} \sim 1 - (1-\beta)\bar{g}$, so we define

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

which is simpler to analyze using existing results.

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

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

with $(\beta, \bar{g}) = \left(1 - c_\beta T^{-\nu}, c_g T^{-\delta}\right)$ for $(\nu, \delta) \in [0,1]^2$. Several cases arise depending on the values of δ, ν. These correspond to a_t exhibiting an exact unit root for $\bar{g} = 0$ or $\beta = 1$, a

near-unit root for $\delta+\nu=1$ (Chan and Wei, 1987, and Phillips 1987), a moderate-unit root for $\delta+\nu\in(0,1)$ (Giraitis and Phillips, 2006, Phillips and Magdalinos, 2007) and a very-near-unit root for $\delta+\nu>1$ (Andrews and Guggenberger, 2007). Under x_t satisfying Assumption B, ξ_T satisfies

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

The case $\delta=\nu=0$ corresponds to a standard weakly stationary process; under Assumption B, the result for $\delta+\nu\in(0,1)$ is found in Magdalinos and Phillips (2009, Lemma 3.1(i)) and that for $\delta+\nu=1$ follows from Stock (1994, Example 4, p. 2754), the magnitude for $\delta+\nu>1$ is dominated by that under the exact unit root case ($\bar{g}=0$ or $\beta=1$), i.e. $\xi_T=O_p\left(T^{1/2}\right)$.

Also $\frac{(1-\beta)\bar{g}}{1-\beta\bar{g}}\asymp(1-\beta)\,\bar{g}$ implies that $S_T^*=\sum_{t=1}^T\beta a_t^*+x_t\overset{p}{\asymp}\sum_{t=1}^T\beta a_t+x_t$. 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:

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

i.e.,

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

Hence

$$\bar{g}\sum_{t=1}^T\xi_t=\frac{1}{(1-\beta)}\left(\sum_{t=1}^T x_t-\xi_T\right)+\bar{g}\xi_T. \tag{23}$$

We start with the case $\nu+\delta<1$, where $\xi_T=o\left(\sum_{t=1}^T x_t\right)$. Expression (23) implies that $\bar{g}\sum_{t=1}^T\xi_t=O_p\left(T^{1/2+\nu}\right)$ and hence

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

If $\nu+\delta=1$, then Stock (1994, example 4, p. 2754) shows that

$$\begin{aligned} T^{-1/2}\left(\sum_{t=1}^T x_t-\xi_T\right)&=T^{-1/2}\sum_{i=1}^T\left(1-\left(1-(1-\beta)\,\bar{g}\right)^{T-i}\right)x_i \\ &=O_p\left(1\right), \end{aligned}$$

It follows that $\sum_{t=1}^T x_t-\xi_T=O\left(T^{1/2}\right)$ and expression (23) implies that $\bar{g}\sum_{t=1}^T\xi_t=O_p\left(T^{1/2+\nu}\right)$. Hence

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

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

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

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 (23):

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

When $\delta < 1$, $3/2 - \delta > 1/2$ 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 $\bar{g} \sum_{t=1}^{T} \xi_{t-1}$:

$$\mathsf{sd}\left(T^{-1/2} S_T^*\right) \asymp T^{1-\delta}.$$

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

A.4 Proof of Theorem 3

In this proof, we omit for notational ease the dependence of β, the spectral densities and autocovariances on T (we hold β and T fixed when referring to Lemma 8). Substitute (6) into (1) to get

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

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

$$(1-\beta)\,y_t + \beta \sum_{j=0}^{t-1} \kappa_j^* y_{t-j} = x_t + \beta\varphi_t.$$

Summing yields

$$\sum_{t=1}^{T} \left((1-\beta) - \beta \sum_{j=0}^{t-1} \kappa_j^*\right) y_{T-t+1} = \sum_{t=1}^{T} (x_t + \beta\varphi_t). \tag{24}$$

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 $\kappa^*(1) = 0$, if there exists $\delta < 1$ such that $\kappa_j \sim c_\kappa j^{\delta-2}$

then $\kappa_j^* \sim -c_\kappa j^{\delta-2}$ and $\sum_{j=0}^{T-1} \kappa_j^* \sim \frac{c_\kappa}{1-\delta} T^{\delta-1}$. Under Assumption A, the previous expressions hold letting $\delta = \delta_\kappa$ when $\delta_\kappa \in (0,1)$; when $\delta_\kappa = 0$, there exists $\delta < 0$ such that $\kappa_j = O\left(j^{\delta-2}\right)$ and $\kappa_j^* = O\left(j^{\delta-2}\right)$ since Assumption A.3 rules out $\kappa_j \sim c_\kappa j^{-2}$.

Let $\beta = 1 - c_\beta T^{-\nu}$. Defining $y_t^- = y_t 1_{\{t \leq 0\}}$, we made the following assumptions about φ_t :

$$\begin{cases} \varphi_t = \kappa(L)\, y_t^-, & \text{if } \delta_\kappa \in \left(\frac{1}{2}, 1\right); \\ \Delta\varphi_t = (1-L)\,\kappa(L)\, y_t^-, & \text{if } \delta_\kappa \in \left(0, \frac{1}{2}\right). \end{cases} \tag{25}$$

so $(1 - \beta\kappa(L))\, y_t = x_t$ if $\delta_\kappa \in (1/2, 1)$ or $(1 - \beta\kappa(L))\, \Delta y_t = \Delta x_t$ if $\delta_\kappa \in (0, 1/2)$. Hence $(1 - \beta\kappa(1))\, E(y_t) = E(x_t)$ or $(1 - \beta\kappa(1))\, E(\Delta y_t) = E(\Delta x_t)$ so the random variables y_t, x_t can be expressed in deviation from their expectations. In other words, we may assume without loss of generality and for ease of exposition that $E(x_t) = 0$ since this does not affect the variances and spectral densities.

Consider the case $\nu > 1 - \delta_\kappa$ so $(1-\beta) / \sum_{j=0}^{T-1} \kappa_j^* \to 0$. This rules out $\delta_\kappa = 0$. 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, with $c_\kappa^{**} > 0$, as $\omega \to 0$

$$f_z(\omega) \sim \frac{f_x(0)}{c_\kappa^{**}} \omega^{-2(1-\delta_\kappa)}. \tag{26}$$

Beran (1994, theorem 2.2 p. 45) shows that (26) implies that

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

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 (25), noting that $\kappa(L)\, y_t^- = (\kappa(L) - 1)\, y_t^-$, expression (24) 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 $\sum_{t=1}^{T} x_t = \sum_{t=1}^{T} (1-\kappa(L)) z_t$,

$$\sum_{t=1}^{T} (1-\kappa(L)) y_t + o_p\left(\sum_{t=1}^{T} (1-\kappa(L)) y_t^+\right) = \sum_{t=1}^{T} x_t$$

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

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

i.e.

$$\sqrt{\mathsf{Var}\left(T^{-1/2}\sum_{t=1}^{T} y_t\right)} \asymp \left(T^{1-\delta_\kappa}\right). \tag{27}$$

Now, if $\delta_\kappa \in (0, 1/2)$, defining $\Delta z_t = [\kappa^*(L)]^{-1} \Delta x_t$, and following the previous steps starting from $(1-\beta\kappa(L))\Delta y_t = \Delta x_t$ 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 $\mathsf{Var}\left(\sum_{t=1}^{T} \Delta z_t\right)$ 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 there exists $c_\gamma \neq 0$ such that $\gamma_{\Delta z}(k) \sim c_\gamma k^{-2\delta_\kappa - 1}$. 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

$$\begin{aligned}
\mathsf{Var}\left(T^{-1}\sum_{t=1}^{T} \Delta z_t\right) &= \gamma_{\Delta z}(0) + 2T^{-1}\sum_{k=1}^{T-1} (T-k)\gamma_{\Delta z}(k) \\
&= \left(\gamma_{\Delta z}(0) + 2\sum_{k=1}^{T-1} \gamma_{\Delta z}(k)\right) - 2T^{-1}\sum_{k=1}^{T-1} k\gamma_{\Delta z}(k) \\
&= -\sum_{k=T}^{\infty} \gamma_{\Delta z}(k) - 2T^{-1}\sum_{k=1}^{T-1} k\gamma_{\Delta z}(k) \\
&\asymp \left(T^{-2\delta_\kappa}\right).
\end{aligned} \tag{28}$$

We now consider the case $\nu \leq 1 - \delta_\kappa$, starting with assuming $\delta_\kappa \neq 0$ so $\nu < 1$. Brillinger (1975, theorem 5.2.1) shows that if the covariances of y_t are summable,

$$\frac{\mathsf{Var}\left(T^{-1}\sum_{t=1}^{T} y_t\right)}{f_y(0)} = (2\pi T)^{-1} \int_{-\pi}^{\pi} \frac{\sin^2(T\omega/2)}{\sin^2(\omega/2)} \frac{f_y(\omega)}{f_y(0)} d\omega, \tag{29}$$

where $f_y(\omega)$ is the spectral density of y_t (the results holds for fixed T, in which case y_t is stationary). 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^+. For any given T and β, Lemma 8 shows that $f_y(\omega)$ is also decreasing in such a neighborhood and $\frac{f_y(\omega)}{f_y(0)}$ is bounded. Both functions in the integrand of (29) 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 (29) 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 (29) 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 $[0,\theta_T]$:

$$\begin{aligned}
&(2\pi T)^{-1}\int_0^{\theta_T}\left(T^2\left(1-\frac{1}{3}\left(T^2-1\right)\omega^2+o\left(T^2\omega^2\right)\right)\right)\left(1-c_V T^\nu\omega^{1-\delta_\kappa}+o\left(T^\nu\omega^{1-\delta_\kappa}\right)\right)d\omega\\
&\quad=\frac{T}{2\pi}\left[\theta_T-\frac{1}{9}\left(T^2-1\right)\theta_T^3-\frac{c}{2-\delta_\kappa}T^\nu\theta_T^{2-\delta_\kappa}+\frac{c_V}{3\left(4-\delta_\kappa\right)}\left(T^2-1\right)T^\nu\theta_T^{4-\delta_\kappa}\right]\\
&\quad=\frac{T}{2\pi}\left[T^{-(1+\varepsilon)}-\frac{T^2-1}{9}T^{-3(1+\varepsilon)}-\frac{c_V}{2-\delta_\kappa}T^{\nu-(2-\delta_\kappa)(1+\varepsilon)}+\frac{c_V\left(T^2-1\right)}{3\left(4-\delta_\kappa\right)}T^{\nu-(4-\delta_\kappa)(1+\varepsilon)}\right]\\
&\quad\sim\frac{1}{2\pi}\left[T^{-\varepsilon}-\frac{1}{9}T^{-3\varepsilon}-\frac{c_V}{2-\delta_\kappa}T^{\nu-(1-\delta_\kappa)-(2-\delta_\kappa)\varepsilon}+\frac{c_V}{3\left(4-\delta_\kappa\right)}T^{\nu-(1-\delta_\kappa)-(4-\delta_\kappa)\varepsilon}\right], \qquad (30)
\end{aligned}$$

where c_V is implicitly defined from Lemma 8. Expression (30) 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 $\frac{\mathsf{Var}\left(T^{-1}\sum_{t=1}^T y_t\right)}{f_y(0)} = O(1)$, with $f_y(0) = (1-\beta)^{-2} f_x(0) \asymp \left(T^{2\nu}\right)$. Hence $\mathsf{Var}\left(T^{-1}\sum_{t=1}^T y_t\right) = O\left(T^{2\nu}\right)$ and

$$\mathsf{Var}\left(T^{-1}\sum_{t=1}^T y_t\right)\asymp\left(T^{2\nu}\right). \qquad (31)$$

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

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

with $\widetilde{\kappa}(1)<\infty$. $\widetilde{\kappa}(L) = (1-L)^{-1}(1-\kappa(L))$ so the roots of $\widetilde{\kappa}$ coincide with the values z such that $\kappa(z)=1$, except at $z=1$ for which $\widetilde{\kappa}(1) = \kappa'(1)>0$ (by L'Hospital's rule and assumption A.3). $\kappa(z)=1$ and $c_\kappa>0$ together imply that the roots of $\widetilde{\kappa}(L)$ lie outside the

unit circle ($\kappa(z) < \kappa(1) = 1$ for $|z| \leq 1$, $z \neq 1$) and the process $\widetilde{x}_t$ defined by $\widetilde{\kappa}(L)\widetilde{x}_t = x_t$ is I(0) with differentiable spectral density at the origin by Assumption B (Stock, 1994, p. 2746). Hence y_t satisfies the near-unit root definition of Phillips (1987):

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

and the result follows from Stock (1994, example 4 p. 2754) since $\widetilde{x}_t$ satisfies his conditions (2.1)-(2.3).

A.5 Proof of Theorem 4

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

A.5.1 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 $\delta_\kappa \in (1/2, 1)$:

$$f_y(\omega) = \frac{f_x(\omega)}{|1 - \beta\kappa(e^{-i\omega})|^2} = \frac{f_x(\omega)}{|1 - \beta + \beta(1 - \kappa(e^{-i\omega}))|^2}, \tag{32}$$

which implies

$$\begin{aligned} &f_y(\omega) \\ &= \frac{f_x(\omega)}{(1-\beta)^2 - 2\beta c_\kappa^*(1-\beta)\omega^{1-\delta_\kappa} + \beta^2 c_\kappa^{**}\omega^{2(1-\delta_\kappa)} + o((1-\beta)\omega^{1-\delta_\kappa}) + o\left(\omega^{2(1-\delta_\kappa)}\right)}. \end{aligned} \tag{33}$$

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 $f_{\Delta y}(\omega_j) \underset{\omega_j \to 0^+}{\sim} \frac{1}{c_\kappa^{**}}\omega_j^{-2\delta_\kappa}$ when $\delta_\kappa \in (0, 1/2)$.

A.5.2 Autocorrelations

The autocovariance function of y_t satisfies

$$\gamma_k = \frac{1}{2\pi}\int_0^{2\pi} f_y(\omega)\, e^{ik\omega} d\omega \tag{34}$$

$$= \frac{1}{2\pi}\int_0^{2\pi} f_y(\omega)\cos(k\omega)\, d\omega, \tag{35}$$

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

$$\frac{1}{2\pi T}\sum_{j=1}^{T} f_y\left(\frac{2\pi j}{T}\right)\cos\frac{2\pi jk}{T} \underset{T\to\infty}{\to} \gamma_y(k). \tag{36}$$

We apply Lemma 6 to expression (36) together with (33).When $\nu > 1-\delta_\kappa$, then $1-\beta = o(\omega_j)$ for all Fourier frequencies ω_j, $j = 1, ..., T$. Expression (33) hence implies that

$$\delta_\kappa \in (1/2,1) : f_y(\omega_j) \sim \frac{f_x(\omega_j)}{(2\pi)^{2(1-\delta_\kappa)} c_\kappa^{**} (j/T)^{2(1-\delta_\kappa)}};$$

$$\delta_\kappa \in (0,1/2) : f_{\Delta y}(\omega_j) \sim \frac{f_x(\omega_j)}{(2\pi)^{-2\delta_\kappa} c_\kappa^{**} (j/T)^{-2\delta_\kappa}}. \tag{37}$$

We refer to Lemma 6 where we let $\delta = 2(1-\delta_\kappa)$ if $\delta_\kappa \in (1/2,1)$ and $\delta = -2\delta_\kappa$ if $\delta_\kappa \in (0,1/2)$. Then for $k = o(T)$:

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

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

A.6 Proof of Theorem 5

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 (33) implies that as $\omega_j \to 0^+$ and for $\nu > 1-\delta_\kappa$,

$$\delta_\kappa \in (1/2,1) : \log f_y(\omega_j) = \log f_x(\omega_j) - \log\left(\beta^2 c_\kappa^{**}\omega_j^{2(1-\delta_\kappa)} + o\left(\omega_j^{2(1-\delta_\kappa)}\right)\right),$$

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

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 (33) implies that:

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

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

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 $h(\omega_j)$,[12] 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

$$\widehat{d} = (1-\delta_\kappa) + \frac{1}{2}\frac{\sum_{j=1}^n \left(\log \phi_T(\omega_j) - \overline{\log \phi_T}\right)\left(h(\omega_j) - \overline{h}\right)}{\sum_{j=1}^n \left(h(\omega_j) - \overline{h}\right)^2} + o_p(1). \tag{38}$$

where as $n \to \infty$,

$$\sum_{j=1}^n \left(h(\omega_j) - \overline{h}\right)^2 \sim 4n. \tag{39}$$

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

$$\tau_T \left[\log \widehat{f}_{y,T}(\omega_j) - \log f_y(\omega_j)\right] \xrightarrow{p} 0, \tag{40}$$

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 (39) 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 $\widehat{d} - (1-\delta_\kappa) \xrightarrow{p} 0$.

A.7 Derivation of models for the forward premium

We derive expression (1) for $y_t = i_t - i_t^*$ from the money-income and Taylor rule models of Engel and West (2005). We show below that both of these models imply a relationship between the log spot exchange rate s_t and y_t of the form

$$s_t = \alpha y_t + b' z_t, \tag{41}$$

where z_t consists of price, money, income, inflation, output gap money demand shock and policy shock differentials, and the real exchange rate, and b is a vector of parameters that is

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

derived below for each model. Substituting in the UIP equation (20) and re-arranging yields

$$\begin{aligned} s_t + y_t &= E_t s_{t+1} - \rho_t \\ (1+\alpha)\, y_t + b' z_t &= \alpha E_t y_{t+1} + b' E_t z_{t+1} - \rho_t \\ y_t &= \frac{\alpha}{1+\alpha} E_t y_{t+1} + \frac{1}{1+\alpha}\left[b' E_t \Delta z_{t+1} - \rho_t\right]. \end{aligned}$$

This is in the form (1) with $\beta = \frac{\alpha}{1+\alpha}$ and $x_t = (1-\beta)\left[b' E_t \Delta z_{t+1} - \rho_t\right]$.

Now, we derive (41) for each of the two models in Engel and West (2005).

Money-income model The money market relationship for the home country (Engel and West, 2005, Equation (4) on p. 492) is given by

$$m_t = p_t + \gamma y_t - \alpha i_t + v_{mt}, \tag{42}$$

where m_t is the log of the home money supply, p_t is the log of the home price level, i_t is the level of the home interest rate, y_t is the log of output, and v_{mt} is a shock to money demand. A similar relationship holds for the foreign country with variables $m_t^*, p_t^*, y_t^*, i_t^*$ and v_{mt}^*, and identical coefficients α and γ. The nominal exchange rate is given by

$$s_t = p_t - p_t^* + q_t \tag{43}$$

where q_t is the (exogenous) real exchange rate (Engel and West, 2005, Equation (5) on p. 493). Subtracting the foreign from the home money market relationship yields

$$p_t - p_t^* = m_t - m_t^* + \gamma\left(y_t^* - y_t\right) + v_{mt}^* - v_{mt} + \alpha\left(i_t - i_t^*\right).$$

Substituting this into (43) yields (41) with $y_t = i_t - i_t^*$ and

$$b' z_t = m_t - m_t^* + \gamma\left(y_t^* - y_t\right) + v_{mt}^* - v_{mt} + q_t.$$

Taylor rule model Suppose the home country follows the Taylor rule (Engel and West, 2005, Equation (9) on p. 494)

$$i_t = \beta_1 y_t^g + \beta_2 \pi_t + v_t, \tag{44}$$

where $\pi_t = p_t - p_{t-1}$ and y_t^g is the "output gap". The foreign country follows the Taylor rule (Engel and West, 2005, Equation (10) on p. 494)

$$i_t^* = -\beta_0\left(s_t - \bar{s}_t^*\right) + \beta_1 y_t^{*g} + \beta_2 \pi_t^* + v_t^*, \tag{45}$$

where $\beta_0 \in (0,1)$ and $\bar{s}_t^*$ is the target for the exchange rate. Assume further that $\bar{s}_t^* = p_t - p_t^*$ (the Purchasing Power Parity level of the exchange rate), see Engel and West (2005, Equation (11) on p. 495). Subtracting (45) from (44) yields

$$i_t - i_t^* = \beta_0 s_t - \beta_0\left(p_t - p_t^*\right) + \beta_1\left(y_t^g - y_t^{*g}\right) + \beta_2\left(\pi_t - \pi_t^*\right) + \left(v_t - v_t^*\right).$$

Re-arranging the above equation yields (41) with $y_t = i_t - i_t^*$, $\alpha = 1/\beta_0$, and

$$b' z_t = (p_t - p_t^*) - \frac{\beta_1}{\beta_0}\left(y_t^g - y_t^{*g}\right) - \frac{\beta_2}{\beta_0}(\pi_t - \pi_t^*) - \frac{1}{\beta_0}(v_t - v_t^*).$$

References

Abadir, K. M., W. Distaso, and L. Giraitis (2007). Nonstationarity-extended local Whittle estimation. *Journal of Econometrics 141*, 1353–1384.

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.

b. Bullard, J. and S. Eusepi (2005). Did the great inflation occur despite policymaker commitment to a Taylor rule? *Review of Economic Dynamics 8*(2), 324 – 359.

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

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.

Benhabib, J. and C. Dave (2014). Learning, large deviations and rare events. *Review of Economic Dynamics 17*(3), 367–382.

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-Rico, V. and J. Gonzalo (2014). Summability of stochastic processes (a generalization of integration and co-integration valid for non-linear processes). *Journal of Econometrics 178*, 331–341.

Bollerslev, T., V. Todorov, and L. Xu (2015). Tail risk premia and return predictability. *Journal of Financial Economics 118*(1), 113–134.

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

Branch, W. A. and G. W. Evans (2011). Learning about risk and return: A simple model of bubbles and crashes. *American Economic Journal: Macroeconomics 3*(3), 159–191.

Branch, W. A. and G. W. Evans (2017). Unstable inflation targets. *Journal of Money, Credit and Banking 49*(4), 767–806.

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.

Bullard, J. B. and S. Eusepi (2005). Did the great inflation occur despite policymaker commitment to a Taylor rule? *Review of Economic Dynamics 8*, 3244–359.

Bullard, J. B., G. W. Evans, and S. Honkapohja (2010). A model of near-rational exuberance. *Macroeconomic Dynamics 14*(2), 166–188.

Campbell, J. Y. (2014). Empirical asset pricing: Eugene fama, lars peter hansen, and robert shiller. *The Scandinavian Journal of Economics 116*(3), 593–634.

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.

Carceles-Poveda, E. and C. Giannitsarou (2008). Asset pricing with adaptive learning. *Review of Economic dynamics 11*(3), 629–651.

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

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. 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.

Chevillon, G. and S. Mavroeidis (2017). Learning can generate long memory. *Journal of Econometrics 198*(1), 1 – 9.

Cho, I.-K. and T. J. Sargent (2008). self-confirming equilibria. In S. N. Durlauf and L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics*. Basingstoke: Palgrave Macmillan.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance 66*(4), 1047–1108.

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

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

Engel, C. (2014). Exchange rates and interest parity. *Handbook of International Economics 4*, 453.

Engel, C. (2016). Exchange rates, interest rates, and the risk premium. *The American Economic Review 106*(2), 436–474.

Engel, C., N. C. Mark, and K. D. West (2008). Exchange rate models are not as bad as you think. In K. R. D. Acemoglu and M. Woodford (Eds.), *NBER Macroeconomics Annual 2007, Volume 22*, pp. 381–441. University of Chicago Press.

Engel, C. and K. D. West (2004). Accounting for exchange-rate variability in present-value models when the discount factor is near 1. *The American Economic Review 94*(2), 119.

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

Eusepi, S. and B. Preston (2011). Expectations, learning, and business cycle fluctuations. *American Economic Review 101*(6), 2844–72.

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

Evans, G. W., S. Honkapohja, T. J. Sargent, and N. Williams (2013). Bayesian model averaging, learning, and model selection. In T. J. Sargent and J. Vilmunenn (Eds.), *Macroeconomics at the Service of Public Policy*, pp. 99–119. Oxford University Press.

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

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

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.

Granger, C. W. J. (1986). Developments in the study of cointegrated economic variables. *Oxford Bulletin of Economics and Statistics 48*(3), 213–228.

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.

Hodges, J. L. and E. L. Lehmann (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics 34*(2), 598–611.

Hommes, C. and G. Sorger (1998). Consistent expectations equilibria. *Macroeconomic Dynamics 2*(03), 287–321.

Johansen, S. (2008). Fractional autoregressive processes. *Econometric Theory 24*, 651–676.

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

Magdalinos, T. and P. C. B. Phillips (2009). Limit theory for cointegrated systems with moderately integrated and moderately explosive regressors. *Econometric Theory 25*, 482–526.

Malmendier, U. and S. Nagel (2016). Learning from inflation experiences. *Quarterly Journal of Economics 131*(1), 53–87.

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.

Orphanides, A. and J. Williams (2004). Imperfect knowledge, inflation expectations, and monetary policy. In *The inflation-targeting debate*, pp. 201–246. University of Chicago Press.

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

Phillips, P. C. B. (2015). Halbert White Jr. Memorial JFEC lecture: Pitfalls and possibilities in predictive regression. *Journal of Financial Econometrics 13*(3), 521–555.

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

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

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

Schotman, P. C., R. Tschernig, and J. Budek (2008). Long memory and the term structure of risk. *Journal of Financial Econometrics 6*(4), 459–495.

Shimotsu, K. (2010). Exact local Whittle estimation of fractional integration with unknown mean and time trend. *Econometric Theory 26*, 501–540.

Shimotsu, K. and P. C. B. Phillips (2005). Exact local Whittle estimation of fractional integration. *The Annals of Statistics 33*(4), 1890–1933.

Slobodyan, S. and R. Wouters (2012). Learning in a medium-scale dsge model with expectations based on small forecasting models. *American Economic Journal: Macroeconomics 4*(2), 65–101.

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.

West, K. D. (2012). Econometric analysis of present value models when the discount factor is near one. *Journal of Econometrics 171*(1), 86–97.