Generalized Local-to-Unity Models*

Liyu Dou and Ulrich K. Müller
CUHK Shenzhen and Princeton University

February 2021

Abstract

We introduce a generalization of the popular local-to-unity model of time series persistence by allowing for p autoregressive roots and $p-1$ moving average roots close to unity. This generalized local-to-unity model, GLTU(p), induces convergence of the suitably scaled time series to a continuous time Gaussian ARMA($p, p-1$) process on the unit interval. Our main theoretical result establishes the richness of this model class, in the sense that it can well approximate a large class of processes with stationary Gaussian limits that are not entirely distinct from the unit root benchmark. We show that Campbell and Yogo's (2006) popular inference method for predictive regressions fails to control size in the GLTU(2) model with empirically plausible parameter values, and we propose a limited-information Bayesian framework for inference in the GLTU(p) model and apply it to quantify the uncertainty about the half-life of deviations from Purchasing Power Parity.

Keywords: Continuous time ARMA process; Convergence; Approximability
JEL Codes: C22; C51

*Müller gratefully acknowledges financial support from the National Science Foundation through grant SES-1627660. Dou gratefully acknowledges financial support from The Chinese University of Hong Kong, Shenzhen through grant UDF01001490. We thank the co-editor, two anonymous referees, David Papell, Mark Watson and participants at the 30th $(EC)^2$ conference for helpful comments and suggestions.

1 Introduction

This paper proposes a flexible asymptotic framework for the modelling of persistent time series. Our starting point is an empirical observation: For many macroeconomic time series, such as the unemployment rate, interest rates, labor's share of national income, real exchange rates, price earnings ratios, etc., tests for an autoregressive unit root are often inconclusive, or rejections are not exceedingly significant. As such, the unit root model is a natural benchmark for empirically plausible persistence modelling. At the same time, most economic models assume that these time series are stationary. What is more, econometric techniques based on an assumption of an exact unit root can yield highly misleading inference under moderate deviations of the unit root model, as demonstrated by Elliott (1998).

These concerns have generated a large literature on econometric modelling and inference with the local-to-unity model.[1] Specifically, a stationary local-to-unity (LTU) model of the scalar time series $x_{T,t}$ is of the form

$$(1 - \rho_T L)(x_{T,t} - \mu) = u_t, \quad t = 1, \ldots, T \tag{1}$$

where L is the lag operator, $\rho_T = 1 - c/T$ for some fixed $c > 0$ and u_t is a mean-zero I(0) disturbance satisfying a functional central limit theorem $T^{-1/2} \sum_{t=1}^{\lceil \cdot T \rceil} u_t \Rightarrow W(\cdot)$ with W a Wiener process with variance equal to the long-run variance of u_t. In this model

$$T^{-1/2}(x_{T,\lceil \cdot T \rceil} - x_{T,1}) \Rightarrow J_1(\cdot) - J_1(0) \tag{2}$$

where J_1 is a stationary Ornstein-Uhlenbeck process with parameter c, the continuous time analogue of an AR(1) process. The process (1) is the local asymptotic alternative of an autoregressive unit root (cf. Elliott, Rothenberg, and Stock (1996), Elliott (1999)). As such, it is impossible to perfectly discriminate between a LTU process and a unit root process, even as $T \to \infty$. Correspondingly, for any finite c, the measure of $J_1(\cdot) - J_1(0)$ is mutually absolutely continuous with respect to the measure of W, the continuous time analogue of a unit root process. LTU asymptotics thus properly reflect the empirical ambivalence of unit root tests noted above.

[1] See Bobkoski (1983), Chan and Wei (1987), Phillips (1987), Stock (1991), Elliott and Stock (1994), Cavanagh, Elliott, and Stock (1995), Wright (2000a), Moon and Phillips (2000), Elliott and Stock (2001), Gospodinov (2004), Valkanov (2003), Torous, Valkanov, and Yan (2004), Rossi (2005), Campbell and Yogo (2006), Jansson and Moreira (2006) and Mikusheva (2007, 2012), among many others.

But the LTU model is clearly not the only persistence model with this feature, even with attention restricted to stationary models. After all, the properties of the limiting process J_1 are governed by a single parameter c, and the long-range dependence of J_1 are those of a continuous time AR(1). For instance, in the LTU model, the correlation between $x_{T,\lceil sT\rceil}$ and $x_{T,\lceil rT\rceil}$ converges to $e^{-c|s-r|}$. It is not clear why this very particular form of long-range dependence should adequately model the persistence properties of macroeconomic time series.

This paper proposes a more flexible asymptotic framework by allowing for p autoregressive roots and $p-1$ moving-average roots local-to-unity for $p \geq 1$, so that

$$(1-\rho_{T,1}L)(1-\rho_{T,2}L)\cdots(1-\rho_{T,p}L)(x_{T,t}-\mu) = (1-\gamma_{T,1}L)\cdots(1-\gamma_{T,p-1}L)u_t,$$

where $\rho_{T,j} = 1-c_j/T$ and $\gamma_{T,j} = 1-g_j/T$ for fixed $\{c_j\}_{j=1}^p$ and $\{g_j\}_{j=1}^{p-1}$ (with some conditions on these parameters as specified in Section 2 below). With $p=1$, this "generalized local-to-unity" model GLTU(p) nests the familiar LTU model (1). A first result of this paper is the convergence of the GLTU(p) model, that is, in analogy to (2),

$$T^{-1/2}(x_{T,\lceil \cdot T\rceil} - x_{T,1}) \Rightarrow J_p(\cdot) - J_p(0) \tag{3}$$

where J_p is a stationary continuous time Gaussian ARMA$(p, p-1)$ process with parameters $\{c_j\}_{j=1}^p$ and $\{g_j\}_{j=1}^{p-1}$.

The GLTU(p) model sets the difference between the number of local-to-unity autoregressive and moving-average parameters to exactly one. This ensures that the limit process J_p still resembles a Wiener process: For instance, if instead $(1-\rho_{T,1}L)(1-\rho_{T,2}L)(x_{T,t}-\mu) = u_t$, the large sample properties of $x_{T,t}$ would be more akin to an I(2) process, with the suitably scaled limit of $x_{T,\lceil \cdot T\rceil} - x_{T,1}$ converging to a limit process that is absolutely continuous with respect to the measure of an *integrated* Wiener process $\int_0^{\cdot} W(r)dr$. In contrast, the measure of $J_p(\cdot) - J_p(0)$ is mutually absolutely continuous with respect to the measure of W, so just as for the LTU model, the GLTU(p) model cannot be perfectly discriminated from the unit root model, even asymptotically.

While clearly more general than the standard local-to-unity model (1), one might still worry about the appropriateness of the GLTU(p) model for generic persistence modelling of macroeconomic time series. Our main theoretical result addresses this concern by establishing the richness of the GLTU(p) model class. Recall that the total variation distance between

two probability measures is the difference in the probability they assign to an event, maximized over all events. We show in Section 3 below that for any given stationary Gaussian limiting process G whose measure of $G(\cdot) - G(0)$ is mutually absolutely continuous with respect to the measure of W, and a mild regularity constraint on the spectral density of G, for any $\varepsilon > 0$ there exists a finite p_ε and GLTU(p_ε) model such that the measure of the induced limiting process J_{p_ε} is within ε of the measure of G in total variation norm. In other words, for small ε, the stochastic properties of J_{p_ε} and G are nearly identical. Thus, positing a GLTU model is nearly without loss of generality for the large sample modelling of persistent stationary processes that cannot be distinguished from a unit root process with certainty in large samples.

In practice, applications of the GLTU model involve the choice of a finite p and the determination of the corresponding $2p-1$ GLTU(p) model parameters $\{c_j\}_{j=1}^{p}$ and $\{g_j\}_{j=1}^{p-1}$. This is perfectly analogous to the modelling of generic covariance stationary processes as finite order AR, MA or ARMA process. The implementation is relatively harder for the GLTU mode, however: As noted above, since the LTU model cannot be perfectly discriminated from the unit root model, the parameter c in (1) cannot be consistently estimated. By the same logic, neither the value of p, nor the parameters $\{c_j\}_{j=1}^{p}$ and $\{g_j\}_{j=1}^{p-1}$ for a given p can be consistently estimated. This impossibility is simply the flip-side of the arguably attractive property of GLTU asymptotics to appropriately capture the empirical ambivalence of unit root tests.

With that in mind, we suggest a limited-information framework for likelihood based inference with the GLTU(p) model. Note that (3) implies, for any fixed integer N

$$\{T^{-1/2}(x_{T,\lceil jT/N \rceil} - x_{T,1})\}_{j=1}^{N} \Rightarrow \{J_p(j/N) - J_p(0)\}_{j=1}^{N}. \tag{4}$$

Thus, with attention restricted to the N observations on the left-hand side of (4), large-sample inference about the GLTU parameters is equivalent to inference given N discretely sampled observations from a continuous time Gaussian ARMA($p, p-1$) process. But this latter problem is well-studied (cf. Phillips (1959), Jones (1981), Bergstrom (1985), Jones and Ackerson (1990), for example), and we show how to obtain a numerically accurate approximation to the likelihood by a straightforward Kalman filter.

We use this framework for two conceptually distinct empirical exercises. First, we show that inference methods derived to be valid in the LTU model can be highly misleading under an empirically plausible GLTU(2) model. In particular, we consider Campbell and Yogo's

(2006) popular test for stock return predictability. By construction, this test controls size in the LTU model. But we find that in the GLTU(2) model, it exhibits severe size distortions, even if the GLTU(2) parameters are restricted to be within a two log-points neighborhood of the peak of the limited-information likelihood for the price-dividend ratio. In other words, unless one has good reasons to impose that the long-range persistence patterns of potential stock price predictors are of the AR(1) type, the Campbell and Yogo (2006) test is not a reliable test of the absence of predictability. This points to Wright's (2000b) test as an attractive alternative that remains robust irrespective of the persistence properties of the predictor.

Second, and more constructively, we conduct limited-information Bayesian inference about the half-life of the US/UK real exchange rate deviations, using the long-span data from Lothian and Taylor (1996). We suggest forming a prior on the GLTU parameters in terms of the smoothness of the implied continuous time ARMA spectral density. The functional form of this spectral density allows for a very compact and easy-to-evaluate expression for one such measure. Substantively, we find that the GLTU model with $p \geq 2$ is strongly preferred by the data as indicated by the corresponding Bayes factors, while at the same time leading to much larger posterior half-life estimates. This illustrates that allowing for the generality provided by the GLTU model can substantially alter conclusions about economic quantities of interest.

The computational convenience of the limited-information likelihood also potentially enables the numerical determination of asymptotically valid frequentist inference in the GLTU(p) model, at least for a given moderately large value of p. In the appendix, we determine location and scale invariant tests of $H_0 : p = 1$ against $H_1 : p > 1$ using the algorithm of Elliott, Müller, and Watson (2015) to deal with the composite nature of the hypotheses. We find that applied to the US/UK real exchange rate data, these tests reject at the 1% level, corroborating the Bayesian finding that the LTU model is unable to adequately capture the low-frequency properties of these series.

This paper contributes to a large literature on alternative models of persistence, such as the fractional model (see Robinson (2003) for an overview), the stochastic local-to-unity model of Lieberman and Phillips (2014, 2017) or the three parameter model of Müller and Watson (2016). These models are generalizations of the unit root model that for almost all parameter values can be perfectly discriminated from this benchmark, at least in large samples, so they do not fall into the class of models this paper focusses on. The scalar GLTU

model is closely connected to the VAR(1) LTU model considered by Phillips (1988), Stock and Watson (1996), Stock (1996) or Phillips (1998): The marginal process for a scalar time series of a VAR(1) LTU model is in the GLTU class, since sums of finite order AR processes are finite order ARMA processes, and as we demonstrate below, the GLTU model can be represented as a weighted average of a latent p-dimensional LTU VAR(1).

Our main theoretical result on the approximability of continuous time Gaussian processes is related in spirit to the approximability of the second order properties of discrete time stationary processes by the finite order ARMA class—see, for instance, Theorem 4.4.3 of Brockwell and Davis (1991) for a textbook exposition. The continuous time case is subtly different, though, since spectral densities are then functions on the entire real line (and not confined to the interval $[-\pi, \pi]$). What is more, we obtain approximability in total variation distance, and not just for a metric on second order properties. We are not aware of any closely related results in the literature.

The remainder of the paper is organized as follows. Section 2 introduces the GLTU(p) model in detail and formally establishes its limiting properties. Section 3 studies the richness of the GLTU(p) model class and contains the main theoretical result. Section 4 develops a straightforward Kalman filter to evaluate the limited-information likelihood. Section 5 contains the two empirical illustrations, and is followed by a concluding Section 6. Proofs are collected in an appendix.

2 The GLTU(p) Model

2.1 Definition

We make the following assumptions about the building blocks of the GLTU(p) model

$$(1-\rho_{T,1}L)(1-\rho_{T,2}L)\cdots(1-\rho_{T,p}L)(x_{T,t}-\mu) = (1-\gamma_{T,1}L)\cdots(1-\gamma_{T,p-1}L)u_t \qquad (5)$$

where $\rho_{T,j} = 1 - c_j/T$ and $\gamma_{T,j} = 1 - g_j/T$.

Condition 1 *(i) The innovations $\{u_t\}_{t=-\infty}^{\infty}$ are mean-zero covariance stationary with absolutely summable autocovariances and satisfy $T^{-1/2}\sum_{t=1}^{\lceil \cdot T \rceil} u_t \Rightarrow W(\cdot)$, where $W(\cdot)$ is a Wiener process of variance ω^2.*

(ii) The parameters $\{c_j\}_{j=1}^{p}$ and $\{g_j\}_{j=1}^{p-1}$ do not depend on T and have positive real parts. They can be complex valued, but if they are, then they appear in conjugate pairs, so that

the polynomials $a(z) = \prod_{j=1}^{p}(c_j + z) = z^p + \sum_{j=1}^{p} a_j z^{p-j}$ *and* $b(z) = \prod_{j=1}^{p-1}(g_j + z) = z^{p-1} + \sum_{j=0}^{p-2} b_j z^j$ *have real coefficients.*

(iii) For all T*, the process* $\{x_{T,t}\}_{t=-\infty}^{\infty}$ *is covariance stationary.*

The high-level Condition 1 (i) allows for flexible weak dependence in the innovations u_t. Part (ii) ensures that a covariance stationary distribution of $x_{T,t}$ exists, and that the limiting continuous time Gaussian ARMA process J_p is causal and invertible. Part (iii) implicitly restricts the initial condition $(x_{T,0}, \ldots, x_{T,-p+1})$ to also be drawn from this covariance stationarity model.

2.2 Continuous Time Gaussian ARMA Processes

Following Brockwell (2001), a mean-zero stationary continuous time Gaussian ARMA$(p, p-1)$ process J_p with parameters $\{c_j\}_{j=1}^{p}$, $\{g_j\}_{j=1}^{p-1}$ and ω^2 of Condition 1 (ii), denoted CARMA$(p, p-1)$ process in the following, can be written as a scalar *observation*

$$J_p(s) = \mathbf{b}'\mathbf{X}(s) \tag{6}$$

of the $p \times 1$ *state* process $\mathbf{X}$ with

$$\mathbf{X}(s) = e^{\mathbf{A}s}\mathbf{X}(0) + \int_0^s e^{\mathbf{A}(s-r)}\mathbf{e}dW(r) \tag{7}$$

where $\mathbf{X}(0) \sim \mathcal{N}(0, \mathbf{\Sigma})$ is independent of the scalar Wiener process W of variance ω^2,

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_p & -a_{p-1} & -a_{p-2} & \cdots & -a_1 \end{pmatrix}, \mathbf{e} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-2} \\ 1 \end{pmatrix}$$

and the coefficients a_j and b_j are defined in Condition 1 (ii). The covariance matrix of $\mathbf{X}(0)$, and hence $\mathbf{X}(s)$, is given by

$$\mathbf{\Sigma} = E[\mathbf{X}(0)\mathbf{X}(0)'] = \omega^2 \int_{-\infty}^{0} e^{-\mathbf{A}r}\mathbf{e}\mathbf{e}'e^{-\mathbf{A}'r}dr, \tag{8}$$

the autocovariance function of J_p is $\gamma_p(r) = E[J_p(s)J_p(s+r)] = \mathbf{b}'e^{\mathbf{A}|r|}\mathbf{\Sigma b}$, and, with $i = \sqrt{-1}$, the spectral density $f_p : \mathbb{R} \mapsto \mathbb{R}$ of J_p satisfies

$$f_{J_p}(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\lambda r}\gamma_p(r)dr = \frac{\omega^2}{2\pi}\frac{|b(i\lambda)|^2}{|a(i\lambda)|^2} = \frac{\omega^2}{2\pi}\frac{\prod_{j=1}^{p-1}(\lambda^2+g_j^2)}{\prod_{j=1}^{p}(\lambda^2+c_j^2)}. \tag{9}$$

Theorem III.17 of Ibragimov and Rozanov (1978) implies that the measure of J_p is mutually absolutely continuous with the measure of J_1 for any fixed $c = c_1 > 0$. Since the measure of $J_1(\cdot) - J(0)$ is mutually absolutely continuous with the measure of W, so the same holds true for $J_p(\cdot) - J_p(0)$.

2.3 Limit Theory

The usual state space representation of the discrete time ARMA process (5) in the definition of the GLTU(p) model is not obviously related to the state space representation (6) and (7) of the CARMA($p, p-1$) process. But it turns out that one can rewrite the former in the form

$$x_{T,t} = \mu + \mathbf{b}'\mathbf{Z}_{T,t} \tag{10}$$

$$\mathbf{Z}_{T,t} = (\mathbf{I} + \mathbf{A}/T)\mathbf{Z}_{T,t-1} + \mathbf{e}u_t \tag{11}$$

where $\mathbf{Z}_{T,t} \in \mathbb{R}^p$, mimicking (6) and (7). This is the key step in the proof of the following theorem, which establishes the large sample relationship between the GLTU(p) model and the corresponding CARMA($p, p-1$) model J_p.

Theorem 1 *Under Condition 1, the GLTU(p) model satisfies* $T^{-1/2}(x_{T,\lceil \cdot T \rceil} - \mu) \Rightarrow J_p(\cdot)$.

Equations (10) and (11) demonstrate that the GLTU model for $x_{T,t}$ amounts to a linear combination of the LTU VAR(1) process $\mathbf{Z}_{T,t}$ with matrix LTU parameter $\mathbf{A}$. As noted in the introduction, the LTU VAR(1) model has been considered in the literature to jointly model several persistent time series. It follows from standard results (see, for instance, Corollary 11.1.1 of Lütkepohl (2005)) that linear combinations of a p dimensional LTU VAR(1) are in the GLTU(p) class, for any (stationary) choice of the matrix LTU parameter. For instance, sums of p independent scalar LTU processes form a particular GLTU(p) model.

We are not aware of previous work that models a scalar series $x_{T,t}$ as a linear combination of a latent LTU VAR(1) process to induce more flexible long-run dynamics. Note that if this

is the objective, then leaving the LTU matrix parameter and the innovation covariance matrix of the latent VAR process unrestricted leads to a severely redundant parameterization. In the non-redundant parameterization of (10) and (11), the LTU matrix parameter $\mathbf{A}$ in (11) has only p free parameters, and the covariance matrix of the innovation $\mathbf{e}u_t$ has only one free parameter.

3 Richness of the GLTU(p) Model Class

In this section we explore the range of large sample persistence patterns that GLTU(p) models can induce. Consider a process $x_{T,t}$, not necessarily a GLTU process, that satisfies $T^{-1/2}(x_{T,\lceil \cdot T \rceil} - \mu) \Rightarrow G(\cdot)$ for some mean-zero stochastic process G on the unit interval. How well can a GLTU(p) model approximate the large sample long-range dynamics of $x_{T,t}$, as characterized by the properties of G?

By Theorem 1, this amounts to studying how well the class of CARMA$(p, p-1)$ processes J_p can approximate the process G. As discussed in the introduction, we focus on processes $x_{T,t}$ that are stationary and that cannot be distinguished from a unit root model with certainty, even asymptotically. Since the limiting process of the unit root model is the Wiener process W, the latter condition amounts to requiring that the measure of $G(\cdot) - G(0)$ is absolutely continuous with respect to the measure of W.

The following theorem shows that the GLTU class can closely approximate all such processes, at least under an additional technical assumption.

Theorem 2 *Let G be a mean-zero continuous time stationary Gaussian process on the unit interval satisfying*

(i) the measure of $G(\cdot) - G(0)$ is absolutely continuous with respect to the measure of W;

(ii) G has a spectral density $f_G : \mathbb{R} \to [0, \infty)$ satisfying $\sup_\lambda (1 + \lambda^2) f_G(\lambda) < \infty$ and $\inf_\lambda (1 + \lambda^2) f_G(\lambda) > 0$.

Then for any $\varepsilon > 0$, there exists a CARMA$(p_\varepsilon, p_\varepsilon - 1)$ process J_{p_ε} such that the total variation distance between the measures of G and J_{p_ε} is smaller than ε.

The conclusion of Theorem 2 is that the entirety of the stochastic properties of G can be well approximated by the stochastic properties for some GLTU(p) limiting process J_p, for a large enough but finite p. In addition to condition (i), which formalizes the assumption of the unit root model as a reasonable statistical benchmark, we require the technical condition

(ii) on the spectral density of G. To shed further light on its nature, note that the spectral density of an Ornstein-Uhlenbeck process J_1 with mean reverting parameter $c = 1$ is given by $f_{J_1}(\lambda) = (2\pi)^{-1}\omega^2/(\lambda^2 + 1)$, so that $\lim_{\lambda\to\infty}(1 + \lambda^2)f_{J_1}(\lambda) = \omega^2/(2\pi)$. In fact, it follows from (9) that $\lim_{\lambda\to\infty}(1+\lambda^2)f_{J_p}(\lambda) = \omega^2/(2\pi)$ for any CARMA$(p, p-1)$ process. Condition (ii) of Theorem 2 only requires that $(1 + \lambda^2)f_G(\lambda)$ is bounded away from zero and infinity uniformly in λ, but not that it converges as $\lambda \to \infty$ (so Theorem 2 covers cases where $\sup_\lambda(1 + \lambda^2)|f_G(\lambda) - f_{J_{p_\varepsilon}}(\lambda)|$ is large, even for small ε). It also immediately follows from Theorem III.17 of Ibragimov and Rozanov (1978) that if $(1+\lambda^q)f_G(\lambda)$ is bounded away from zero and infinity uniformly in λ for some $q > 1$, then for any $q \neq 2$, the measure of G is orthogonal to the measure of J_1, and hence the first assumption in Theorem 2 is violated. Thus, given assumption (i), assumption (ii) is arguably fairly mild.

The proof of Theorem 2 is involved. We leverage classic results on the mutual absolute continuity (but not approximability) of Gaussian measures by Ibragimov and Rozanov (1978) to obtain a bound on the entropy norm between the measures of a countable set of characterizing random variables $\{\psi_j(G)\}_{j=1}^\infty$ and those of potential approximating process in terms of their spectral densities, and then apply a locally compact version of the Stone-Weierstrass theorem to uniformly approximate f_G by some $f_{J_{p_\varepsilon}}$. See the appendix for details.

4 A Limited-Information Likelihood Framework

In this section we suggest a framework for conducting large sample inference with the GLTU model. A natural place to start would be the likelihood of J_p. Pham-Dinh (1977) derives the likelihood but notes that it is "too complicated for practical use" (page 390). What is more, it wouldn't be appropriate to treat $T^{-1/2}(x_{T,\lfloor\cdot T\rfloor} - \mu)$ as a realization of $J_p(\cdot)$ directly, since Theorem 1 only establishes weak convergence. To make further progress, we restrict attention to inference that is a function only of the N random variables $\{x_{T,\lceil jT/N\rceil}\}_{j=1}^N$, for some given finite integer N.

4.1 Large Sample Approximation

The following result is immediate from Theorem 1 and the continuous mapping theorem.

Corollary 1 *Under Condition 1, for any fixed integer $N \geq 1$,*

$$\{T^{-1/2}(x_{T,\lceil jT/N\rceil} - \mu)\}_{j=1}^N \Rightarrow \{J_p(j/N)\}_{j=1}^N. \tag{12}$$

An asymptotically justified limited-information likelihood of the GLTU(p) model is thus given by the likelihood of a discretely sampled CARMA($p, p-1$) process. The number N determines the resolution of the limited-information "lens" through which we view the original data $\{x_{T,t}\}_{t=1}^{T}$. The convergence in Theorem 1, and thus in (12), are approximations that show that under a wide range of weak dependence of u_t, Central Limit Theorem type effects yield large sample Gaussianity and a dependence structure that is completely dominated by the long-run dependence properties of the GLTU(p) model. In finite samples, a large N takes these approximations seriously even on a relatively fine grid, so in general, a large N reduces the robustness of the resulting inference. At the same time, a small N leads to a fairly uninformative limited-information likelihood.[2] The choice of N thus amounts to a classic efficiency vs. robustness trade-off. In our applications, we set $N = 50$.

4.2 Numerical Approximation to Limited-Information Likelihood

As noted in the introduction, there are a number of suggestions in the literature on how to obtain the likelihood of a discretely sampled CARMA($p, p-1$) process. One potential difficulty is the computation of covariance matrices involving matrix exponentials (cf. (8)). If the local-to-unity AR roots are distinct, then the companion matrix $\mathbf{A}$ is diagonalizable, so one can rotate the system by the matrix of eigenvectors to avoid this difficulty. But in general, this yields a complex valued system, which requires additional care. What is more, one might not want to rule out a pair of identical local-to-unity AR roots a priori.

We now develop an alternative approach for the computation of the likelihood of $\{J_p(j/N)\}_{j=1}^{N}$ that avoids these difficulties. To this end, consider the discrete time Gaussian ARMA($p, p-1$) process

$$(1-\rho_{T_0,1}L)\cdots(1-\rho_{T_0,p}L)(x^0_{T_0,t}-\mu) = (1-\gamma_{T_0,1}L)\cdots(1-\gamma_{T_0,p-1}L)u^0_t \tag{13}$$

for $t = 1, \ldots, T_0$, where T_0 is large, $u^0_t \sim iid\mathcal{N}(0,\omega^2)$ and $\rho_{T_0,j}$, $\gamma_{T_0,j}$ are defined below (5). As in (10) and (11), $x^0_{T_0,t}$ has the state space representation

$$x^0_{T_0,t} = \mathbf{b}'\mathbf{Z}^0_{T_0,t} + \mu \tag{14}$$

$$\mathbf{Z}^0_{T_0,t} = (\mathbf{I}_p + \mathbf{A}/T_0)\mathbf{Z}^0_{T_0,t-1} + \mathbf{e}u^0_t \tag{15}$$

[2] The number of observations N thus plays a similar role to the number of cosine regression coefficients q in the low-frequency extraction approach of Müller and Watson (2017); in combination with the Kalman filter described below, the approach based on (12) is computationally much more advantageous, however.

with $\mathbf{\Omega}^0_{T_0} = E[\mathbf{Z}^0_{T_0,0}(\mathbf{Z}^0_{T_0,0})']$ satisfying $\operatorname{vec}\mathbf{\Omega}^0_{T_0} = \omega^2(\mathbf{I}_{p^2} - (\mathbf{I}_p + \mathbf{A}/T_0)\otimes(\mathbf{I}_p + \mathbf{A}/T_0))^{-1}\operatorname{vec}(\mathbf{ee}')$.

With $T = T_0$ and $u_t = u^0_t$, the model (13) clearly satisfies Condition 1, so by Corollary 1,

$$\{T_0^{-1/2}(x^0_{T_0,\lceil jT_0/N\rceil} - \mu)\}_{j=1}^N \Rightarrow \{J_p(j/N)\}_{j=1}^N \tag{16}$$

as $T_0 \to \infty$. Furthermore, since $\{x^0_{T_0,t}\}_{t=1}^{T_0}$ is a Gaussian process, this further implies convergence of the corresponding first two moments. Thus, the Gaussian likelihood of $\{T_0^{-1/2}(x^0_{T_0,\lceil jT_0/N\rceil} - \mu)\}_{j=1}^N$ approximates the likelihood of $\{J_p(j/N)\}_{j=1}^N$ arbitrarily well as $T_0 \to \infty$. An accurate Euler-type approximation to the asymptotically justified limited-information likelihood for the $2p+1$ parameters μ, ω^2, $\{c_j\}_{j=1}^p$ and $\{g_j\}_{j=1}^{p-1}$ of the GLTU(p) model can therefore be obtained from a straightforward application of the Kalman filter with state (15) and observations $x^0_{T_0,\lceil jT_0/N\rceil} = x_{T,\lceil jT/N\rceil}$, $j = 1,\ldots,N$, with all other observations of $x^0_{T_0,t}$ treated as missing. In our applications, we found that setting $T_0 = 1000$ leads to results that remain numerically stable also for larger values of T_0.

4.3 Parameterization of GLTU Parameters

A remaining difficulty is the restriction on the parameters $\{c_j\}_{j=1}^p$ and $\{g_j\}_{j=1}^{p-1}$ of Condition 1 (ii). Here we follow Jones (1981), who noted that under Condition 1 (ii), one can rewrite $a(z)$ and $b(z)$ as a product of quadratic factors (and a linear factor if p is odd), where each quadratic factor collapses a potentially conjugate pair of roots into a quadratic polynomial with positive coefficients. For instance, if $c_1 = c_1^r + c_1^i i$ and $c_2 = c_1^r - c_1^i i$ with $c_1^r > 0$ and $c_1^i \in \mathbb{R}$, then $(z+c_1)(z+c_2) = (c_1^i)^2 + (c_1^r)^2 + 2c_1^r z + z^2$, and if c_1 and c_2 are real and positive, $(z+c_1)(z+c_2) = c_1c_2 + (c_1+c_2)z + z^2$. Either way, the resulting quadratic polynomial is of the form $h_1^2 + 2h_2 z + z^2$, with $h_1, h_2 > 0$, and in this parameterization $c_{1,2} = h_2 \pm \sqrt{h_2^2 - h_1^2}$. The same argument applies to the MA polynomial. Thus, we can parameterize $\{c_j\}_{j=1}^p$ and $\{g_j\}_{j=1}^{p-1}$ in terms of the vector $\mathbf{h} = (h_1^c,\ldots,h_p^c,h_1^g,\ldots,h_{p-1}^g)' \in [0,\infty)^{2p-1}$, where for p odd, $c_p = h_p^c$, and for p even, $g_{p-1} = h_{p-1}^g$.

Note that the function $\lambda^2 \mapsto (\lambda^2 + c_1^2)(\lambda^2 + c_2^2)$ potentially has a local minimum at $\lambda = \sqrt{-(c_1^2+c_2^2)/2} = \sqrt{h_1^2 - 2h_2^2} \le h_1$. At the same time, the Nyquist frequency of N discrete limited-information observations is $N\pi$. Thus restricting each value of h to the interval $[0, N\pi]$ still provides full flexibility of the rational spectral density function (9) over the relevant frequency band, and we will impose this restriction in the following.

5 Applications

This section describes two applications of the GLTU(p) model and the limited-information framework of the last section. The first application considers the popular Campbell and Yogo (2006) hypothesis test of no predictability in the presence of a persistent predictor. This test is derived under the assumption that the persistence is of the LTU form, and we consider its size when in fact the persistence is generated by a GLTU(2) model. We find that empirically plausible values of the GLTU(2) parameters for the U.S. price-dividend log-ratio from 1926:12-2018:5 induce large size distortions.

Our second application concerns the quantification of mean reversion in real exchange rates predicted by the theory of purchasing power parity, applied to the long-span data assembled by Lothian and Taylor (1996). We conduct Bayesian inference about the degree of mean reversion in the GLTU(p) model for $p = 1, 2, \ldots, 5$. We find that the results for $p > 1$ are quite different from those for the $p = 1$ LTU model, and that the LTU model provides a substantially worse fit.

Both applications thus show that the standard approach of modelling persistent time series with the LTU model yields potentially misleading empirical conclusions, suggesting a need for the greater flexibility provided by the GLTU(p) model.

5.1 Predictive Regression with a Persistent Predictor

Let $y_{T,t}$ denote the excess stock return in period t, and let $x_{T,t-1}$ denote a persistent potential predictor variable observed at $t-1$, such as the price-dividend log-ratio. A standard LTU formulation of this set-up is

$$y_{T,t} = \mu_y + \beta x_{T,t-1} + e_t, \tag{17}$$

$$(1 - \rho_T L)(x_{T,t} - \mu) = u_t \tag{18}$$

where $\rho_T = 1 - c/T$ for fixed $c > 0$, and the mean-zero disturbances (e_t, u_t) are weakly dependent with correlation r_{eu}. The null hypothesis of no predictability amounts to $H_0 : \beta = 0$, against the alternative that $H_1 : \beta \neq 0$.

As is well understood, the OLS estimator of β in this set-up is biased under $r_{eu} \neq 0$, invalidating standard inference based on the t-statistic on β in (17). While the bias is a function of c, and hence not consistently estimable, several approaches have been devised to obtain valid inference: see, for instance, Elliott and Stock (1994), Cavanagh, Elliott, and Stock (1995),

Campbell and Yogo (2006), Jansson and Moreira (2006) and Elliott, Müller, and Watson (2015). These tests have local asymptotic power against alternatives of the form $\beta = b/T$, $b \neq 0$, but their construction is predicated on the LTU form (18) of predictor persistence. Magdalinos and Phillips (2009) and Kostakis, Magdalinos, and Stamatogiannis (2015) devise an approach that focusses on higher frequency variability of the predictor, which recovers standard normal null distributions for test statistics. This presumably provides robustness also under alternative forms of predictor persistence, although at the cost of no asymptotic local power in the $\beta = b/T$ neighborhood of the null hypothesis.[3] Finally, Wright (2000b) suggested an ingenious approach that does not require any assumptions about the properties of $x_{T,t}$ for its validity (also see Lanne (2002)): Under the null hypothesis of $H_0 : \beta = 0$ in (17), $y_{T,t}$ recovers the true prediction errors e_t up to a constant, so a test of stationarity of $y_{T,t}$ does not overreject irrespective of the properties of the predictor $x_{T,t}$. Furthermore, adapting the arguments leading to Theorem 4 of Wright (2000b) shows that such a test has nontrivial asymptotic power under local alternatives with $\beta = \beta_0 + b/T$ and $x_{T,t}$ satisfying $T^{-1/2}(x_{T,\lceil \cdot T \rceil} - \mu) \Rightarrow G(\cdot)$ as long as G is an almost sure nonzero continuous function.

The most popular approach in practice is Campbell and Yogo (2006)'s test, which corrects for the bias by forming a confidence interval for c, and a Bonferroni-type correction to the critical value. This construction crucially exploits the LTU model (18) for the predictor.[4] At the same time, it is not obvious whether empirically plausible alternative forms of persistence can induce (large) size distortions. Specifically, in contrast to Campbell and Yogo's assumption, suppose $x_{T,t}$ follows a GLTU(2) model

$$(1 - \rho_{T,1}L)(1 - \rho_{T,2}L)(x_{T,t} - \mu) = \left(1 - \gamma_{T,1}L\right) u_t. \tag{19}$$

Does the 10% level Campbell and Yogo (2006) test continue to reject a true null hypothesis of no predictability $H_0 : \beta = 0$ at most 10% of the time under such an alternative form of persistence?

We investigate this issue in the context of the empirical example in Campbell and Yogo (2006), where $y_{T,t}$ is the monthly excess return on the NYSE/AMSE value-weighted monthly index, and $x_{T,t}$ is the corresponding price-dividend log-ratio, averaged over the preceding 12

[3] Similarly, Kasparis, Andreou, and Phillips (2015) develop a robust approach to nonparametric predictive regressions with potentially persistent regressors with local asymptotic power against larger alternatives.

[4] Technically, Campbell and Yogo (2006) assume a non-stationary LTU model with zero initial condition. We therefore impose a zero initial condition in the empirical analysis both for the LTU and the GLTU(2) model.

Figure 1: CRSP Price-Dividend Ratio and Empirically Plausible Limiting Log-Spectra

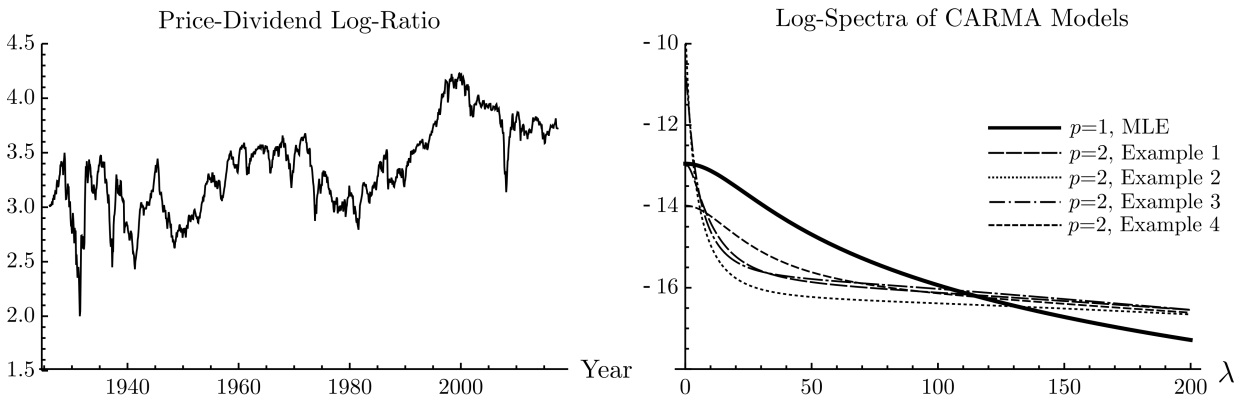


Table 1: Four GLTU(2) Parameters and Resulting Null Rejection Probability of the 10% Level Campbell-Yogo (2006) Test

Example No.	1	2	3	4	LTU-MLE
Value of c_1	5.0	0.7	1.3	17.2	23.1
Value of c_2	230.1	308.1	191.6	264.1	NA
Value of g_1	22.8	17.2	15.5	52.7	NA
Null rejection probability	48.7%	73.7%	46.4%	49.3%	5.8%

months. We updated the Campbell and Yogo (2006) data set to 1098 monthly observations from 1926:12-2018:5 from the database of the Center for Research in Security Prices (CRSP). The left panel in Figure 1 plots $x_{T,t}$.

To obtain empirically plausible parameters of the GLTU(2) model, we first maximize the limited-information likelihood with $N = 50$ as described in Section 4 in the LTU model, yielding the MLE for c equal to 23.1. Call values of $\{c_1, c_2, g_1\}$ "empirically plausible" for the GLTU(2) model (19) if the profiled value over μ and ω^2 of the limited-information likelihood is within two log-points of the LTU maximum likelihood. This definition ensures that a GLTU(2) model with empirically plausible parameter values cannot be distinguished from the baseline LTU model with much confidence.

We then compute the rejection probability of Campbell and Yogo's (2006) nominal 10% level two-sided test of no predictability for data generated from such empirically plausible GLTU(2) processes with $T = 1098$, $\beta = 0$, $(e_t, u_t)'$ i.i.d. mean-zero normal and correlation equal to $r_{eu} = -0.951$, which is the value of r_{eu} estimated by Campbell and Yogo's procedure under the LTU model assumption. (The test is invariant to translation shifts and scale transformations of $y_{T,t}$ and $x_{T,t}$, so the variances of e_t and u_t, as well as the means μ and μ_y are immaterial.) In Table 1 we report the parameter values for four fairly distinct empirically plausible GLTU(2) parameters that induce severe size distortions.[5] The right panel in Figure 1 plots the corresponding log spectral densities of the limiting CARMA(2,1) model, along with the limiting Ornstein-Uhlenbeck process with $c = 23.1$, that is at the limited-information MLE.

We conclude from this exercise that the validity of the Campbell and Yogo (2006) test very much depends on the untestable assumption that the predictor persistence is of the LTU form. Since this is arguably an unattractive assumption, a more compelling test of no predictability is Wright's (2000b) approach, which is asymptotically valid with nontrivial power in the $\beta = b/T$ neighborhood for the entire GLTU class.

5.2 Persistence of Deviations from Purchasing Power Parity

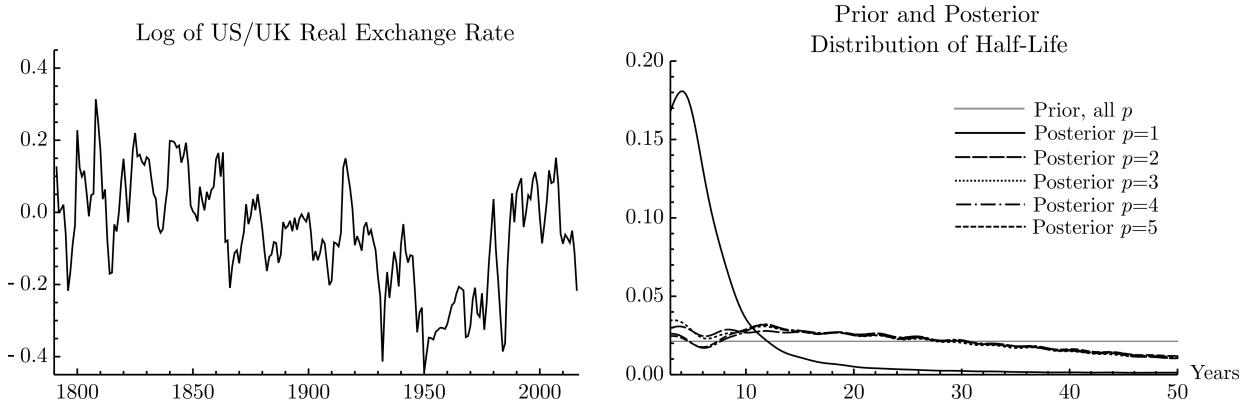
Lothian and Taylor (1996) assembled long-term data on the log US/UK real exchange rate from 1791 to 1990 and estimated half-life deviations of approximately 6 years based on an AR(1) specification. We consider the same data extended through 2016, $x_{T,t}$, and plotted in the left panel of Figure 2.[6] We are interested in quantifying for how long deviations from purchasing power parity persist assuming that the exchange rate $x_{T,t}$ follows a GLTU(p) model.

The traditional definition of the half-life is based on the impulse response of the Wold innovation to $x_{T,t}$, which in general depends not only on the GLTU(p) parameters $\{c_j\}$ and $\{g_j\}$, but also on the short-run dynamics of u_t. See, for instance, Andrews and Chen (1994), Murray and Papell (2002) or Rossi (2005). At the same time, as discussed in Taylor's (2003)

[5] This adds to the analysis by Phillips (2014) and Kostakis, Magdalinos, and Stamatogiannis (2015), who document size distortions of the Campbell and Yogo (2006) test with an AR(1) predictor that exhibits less than LTU persistence.

[6] The extension is based on the FRED series DEXUSUK, SWPPPI and WPSFD49207 for recent values of the exchange rate, and UK and US producer price indices.

Figure 2: Bayesian Limited-Information Analysis of US/UK Real Exchange Rates



survey, the literature on real exchange rates emphasizes mean reversion *in the long run*, and often applies corresponding augmented Dickey-Fuller regressions, which in the context of the LTU model amount to inference about c (also see Murray and Papell (2005) and Stock (1991)).

Impulse responses are most meaningful in the context of a structural model, where innovations are given an explicit interpretation. But the structural interpretation of Wold innovations to the real exchange rates is not obvious. We therefore define the half-life in terms of the following thought experiment: Given the model parameters, suppose we learn that the value of the stationary process $x_{T,t}$ at the time $t = 0$ is one unconditional standard deviation above its mean, but we don't observe any other values of $x_{T,t}$. What is the smallest horizon τ such that the best linear predictor of $x_{T,t}$ given $x_{T,0}$ is within 1/2 unconditional standard deviations of its mean, for all $t \geq \tau$?

The best linear predictor of $x_{T,t}$ given $x_{T,0}$ is proportional to the correlation between $x_{T,0}$ and $x_{T,t}$. Assuming that u_t has more than two moments, Theorem 1 implies that

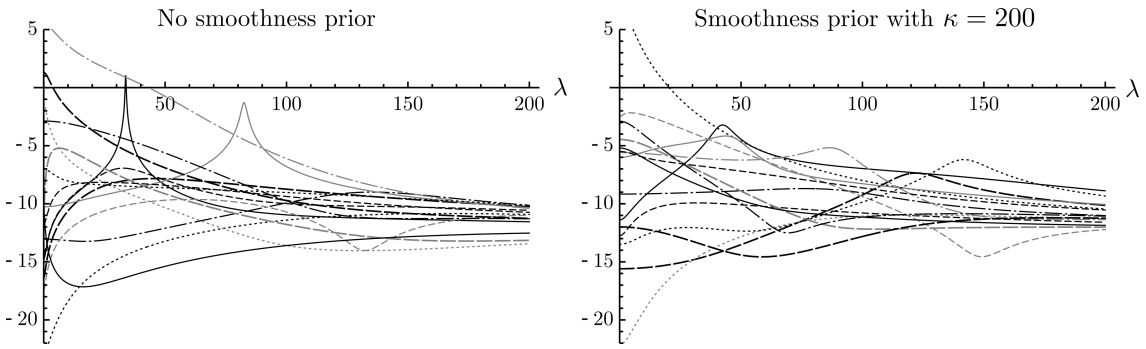
$$T^{-1}E[(x_{T,0} - \mu)(x_{T,\lceil sT \rceil} - \mu)] \to E[J_p(0)J_p(s)] = \mathbf{b}' e^{\mathbf{A}|s|}\mathbf{\Sigma b}$$

so that we obtain the large sample approximation

$$\tau \approx T \inf_r \left\{ r : \left| \frac{\mathbf{b}' e^{\mathbf{A}|s|}\mathbf{\Sigma b}}{\mathbf{b}'\mathbf{\Sigma b}} \right| \leq 1/2 \text{ for all } s \geq r \right\}. \tag{20}$$

For $p = 1$, that is in the LTU model, this definition of a half-life is equivalent to the half-

Figure 3: Random Log-Spectral Density Draws from the Baseline Prior



life of the impulse response relative to the "long-run" shock u_t, which in large samples becomes the impulse response function of J_1. But for $p > 1$, this equivalence breaks down, since the impulse response function of J_p is equal to $\mathbf{1}[s \geq 0]\mathbf{b}'e^{\mathbf{A}s}\mathbf{e}$ (cf. (6) and (7)), while the autocovariance function is $\mathbf{b}'e^{\mathbf{A}|s|}\boldsymbol{\Sigma}\mathbf{b}$. We explore this and other possible alternative definitions of the half-life in the appendix.

In order to avoid evaluating the matrix exponential in (20), note that $\mathbf{b}'e^{\mathbf{A}|s|}\boldsymbol{\Sigma}\mathbf{b}$ can be arbitrarily well approximated by the autocovariance function of the discrete stationary state space system (14) and (15) as $T_0 \to \infty$, so that

$$\tau \approx T \inf_r \left\{ r : \left| \frac{\mathbf{b}'(\mathbf{I}_p + \mathbf{A}/T_0)^{\lceil sT_0 \rceil} \boldsymbol{\Omega}^0_{T_0} \mathbf{b}}{\mathbf{b}'\boldsymbol{\Omega}^0_{T_0}\mathbf{b}} \right| \leq 1/2 \text{ for all } s \geq r \right\}. \tag{21}$$

We again find that choosing $T_0 = 1000$ generates numerically stable results.

We consider the GLTU model with $p = 1, 2, \cdots, 5$, and conduct inference based on the limited-information likelihood for $N = 50$ as discussed in Section 4. We choose the usual improper uninformative priors for the location and scale parameters μ and ω^2. For $\{c_j\}_{j=1}^p$ and $\{g_j\}_{j=1}^{p-1}$, we employ the h parameterization of Section 4, collected in the vector $\mathbf{h} \in [0, 50\pi]^{2p-1}$.

A prior on $\mathbf{h}$ may be obtained by considering its implication for the smoothness of the resulting spectral density. We focus on the second derivative of the log-spectrum as a measure of this smoothness. A calculation detailed in the appendix shows that the special form (9) of the CARMA spectrum f_{J_p} leads to the computationally convenient expression

$$\frac{1}{8\pi}\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\lambda^2}\ln f_{J_p}(\lambda)\right)^2 d\lambda = \sum_{k,j=1}^{p}\frac{1}{(c_k+c_j)^3} + \sum_{k,j=1}^{p-1}\frac{1}{(g_k+g_j)^3} - 2\sum_{k=1}^{p}\sum_{j=1}^{p-1}\frac{1}{(c_k+g_j)^3}. \tag{22}$$

Table 2: Bayesian Limited-Information Analysis of US/UK Real Exchange Rates

	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$
posterior median half-life	5.9	21.6	21.2	22.3	22.3
90% posterior interval	(3.5; 23.5)	(4.6; 45.4)	(4.3; 45.5)	(4.9; 45.8)	(5.1; 45.9)
Bayes factor relative to $p=1$	1.0	36.5	27.0	33.9	27.3

Denote the right hand side of (22) by $\psi(\mathbf{h})$, with $\{c_j\}_{j=1}^{p}$ and $\{g_j\}_{j=1}^{p-1}$ considered functions of $\mathbf{h}$. Our baseline prior on $\mathbf{h} \in [0, 50\pi]^{2p-1}$ is then of the form $\pi_b(\mathbf{h}) \propto \exp[-\kappa\psi(\mathbf{h})]$ for some $\kappa \geq 0$. Roughly speaking, the overall persistence of J_p is determined by the rate at which $f_{J_p}(\lambda)$ declines. The prior π_b is agnostic about this rate. Rather, it penalizes the amount of variation around the average decline. Sufficiently large values of $\kappa > 0$ thus ensure that even with large p, the prior strongly favors spectra of overall smooth shape. Visual inspection of prior draws, as illustrated in Figure 3 for $p = 5$, lead us to choose $\kappa = 200$ for our baseline empirical specification. Additional calculations detailed in the appendix show that the following results are not sensitive to this choice.

The main objective of our empirical illustration is to demonstrate that the GLTU model can yield substantially different empirical results about the half-life. In order to isolate this effect, we adjust the baseline smoothness prior π_b to ensure that the implied prior for the half-life does not mechanically depend on p. Let $\tau(\mathbf{h})$ be the half-life in (20) implied by a given value of $\mathbf{h}$. Then we employ the prior $\pi(\mathbf{h}) \propto \pi_b(\mathbf{h})\pi_\tau(\tau(\mathbf{h}))$, where the function $\pi_\tau : \mathbb{R} \mapsto [0, \infty)$ is such that the prior distribution of $\tau(\mathbf{h})$, measured in years, is uniform on the interval $[3, 50]$ under π.

The posterior is obtained from a random walk Metropolis-Hastings algorithm after analytically integrating out (μ, ω^2). With the Kalman filter approximation to the limited-information likelihood of Section 4, the corresponding half-life approximation (21), and the expression (22), evaluation of the posterior density is very fast. We provide additional computational details in the appendix.

The second and third rows of Table 2 provide summary statistics for the posterior half-life for $1 \leq p \leq 5$, and the right panel of Figure 2 plots the posterior densities. For $p = 1$, the posterior for the half-life is unimodal with a mode of around 4.0 years and a median of 5.9, more or less in line with the original results of Lothian and Taylor (1996). But letting $p > 1$ leads to posteriors with much more mass at substantially longer half-lives.

This accords qualitatively with Murray and Papell's (2005) finding of longer half-life point estimates when allowing for many lags in the autoregression, although their half-lives are computed from impulse responses or sums of autoregressive coefficients, and are thus not directly comparable.

Remarkably, the posterior densities in Figure 3 for $p \geq 2$ are very similar to each other. It seems that once the model is flexible enough, the implications settle, with a wide posterior distribution for the half-life with a median around 22 years.

The Bayes factors relative to the LTU ($p = 1$) model in Table 2 indicate a strong preference by the data for values of $p > 1$, and are quite similar for the values of $p > 1$. In the appendix we derive frequentist tests of $H_0 : p = 1$ against $H_1 : p > 1$. These tests also reject at the 1% level on the US/UK real exchange rate data with $N = 50$.

Overall, these results suggest that the LTU model does not adequately account for the long-run properties of this data, and that accounting for them in a more flexible manner yields substantially longer half-lives of PPP deviations.

6 Conclusion

This paper introduces the GLTU(p) model as a natural generalization of the popular local-to-unity approach to modelling stationary time series persistence. The main theoretical result concerns the richness of this model class: The asymptotic properties of a large class of persistent processes that is not entirely distinct from an I(1) benchmark can be well approximated by some GLTU(p) model.

We further suggest a straightforward approximation to the limited-information asymptotic likelihood of the GLTU(p) model, and derive a computationally convenient prior for the GLTU parameters that penalizes non-smooth spectral densities. The resulting limited-information Bayesian analysis is straightforward to implement and, for p large, flexibly adapts to a wide range of potential low-frequency behavior. The GLTU(p) model thus seems a convenient starting point for the modelling of persistent time series in macroeconomics and finance.

A Appendix

A.1 Proof of Theorem 1

We first show that $x_{T,t}$ has representation (10) and (11). Set $\prod_{j=1}^{p}(z-\rho_{T,j}) = z^p + \sum_{j=1}^{p}\phi_{T,j}z^{p-j}$ and $\prod_{j=1}^{p-1}(z-\gamma_{T,j}) = z^{p-1}+\sum_{j=0}^{p-2}\theta_{T,j}z^j$. The usual state-space representation of the ARMA$(p,p-1)$ process $x_{T,t}$ with innovations u_t is

$$x_{T,t} = \boldsymbol{\theta}_T'\mathbf{V}_{T,t} + \mu \tag{23}$$

$$\mathbf{V}_{T,t} = \boldsymbol{\Phi}_T\mathbf{V}_{T,t-1} + \mathbf{e}u_t \tag{24}$$

where

$$\mathbf{V}_{T,t} = \begin{pmatrix} v_{T,t-p+1} \\ v_{T,t-p+2} \\ \vdots \\ v_{T,t-1} \\ v_{T,t} \end{pmatrix}, \boldsymbol{\Phi}_T = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\phi_{T,p} & -\phi_{T,p-1} & -\phi_{T,p-2} & \cdots & -\phi_{T,1} \end{pmatrix}, \boldsymbol{\theta}_T = \begin{pmatrix} \theta_{T,0} \\ \theta_{T,1} \\ \vdots \\ \theta_{T,p-2} \\ 1 \end{pmatrix}.$$

Let $\mathbf{c} = (c_1,\ldots,c_p)'$ and $\mathbf{g} = (g_1,\ldots,g_{p-1})'$ with elements ordered ascendingly by the real parts, and define the corresponding vectors $\boldsymbol{\rho}_T = (\rho_{T,1},\ldots,\rho_{T,p})'$ and $\boldsymbol{\gamma}_T = (\gamma_{T,1},\ldots,\gamma_{T,p-1})'$.

For any $\mathbf{a} = (a_1,\ldots,a_k)' \in \mathbb{C}^k$, $k \in \{p-1,p\}$ with $\mathrm{Re}(a_j) \le \mathrm{Re}(a_{j+1})$, let $\mathbf{J}(\mathbf{a})$ be the $k\times k$ Jordan matrix with Jordan blocks corresponding to common values of a_j, and let $\mathbf{Q}(\mathbf{a})$ be the $p\times k$ matrix where the m columns of $\mathbf{Q}(\mathbf{a})$ corresponding to the value a of multiplicity m contain the values $\frac{d^l}{dz^l}z^{j-1}|_{z=a}/l!$, $j=1,\ldots,p$, $l=0,\ldots,m-1$. Since $\boldsymbol{\Phi}_T$ and $\mathbf{A}$ are companion matrices, and the roots of $z^p+\sum_{j=1}^{p}\phi_{T,j}z^{p-j}$ and $a(z)$ are $\boldsymbol{\rho}_T$ and $-\mathbf{c}$, respectively, they allow the Jordan decomposition (cf. Brand (1964))

$$\boldsymbol{\Phi}_T = \mathbf{Q}(\boldsymbol{\rho}_T)\mathbf{J}(\boldsymbol{\rho}_T)\mathbf{Q}(\boldsymbol{\rho}_T)^{-1} \tag{25}$$

$$\mathbf{A} = \mathbf{Q}(-\mathbf{c})\mathbf{J}(-\mathbf{c})\mathbf{Q}(-\mathbf{c})^{-1}. \tag{26}$$

From (26), we also have

$$\mathbf{I}+\mathbf{A}/T = \mathbf{Q}(-\mathbf{c})(\mathbf{I}+\mathbf{J}(-\mathbf{c})/T)\mathbf{Q}(-\mathbf{c})^{-1}. \tag{27}$$

Let $\mathbf{F}$ be the $p\times p$ lower triangular Pascal matrix, that is, the first j entries in row j of $\mathbf{F}$ contain the jth binomial coefficients, and let $\mathbf{D}_T = \mathrm{diag}(1,T^{-1},\ldots,T^{1-p})$. Further, let $\mathbf{D}_T^c$

be a diagonal matrix where the diagonal elements corresponding to a Jordan block of $\mathbf{A}$ of size m are equal to $1, T, \ldots, T^{m-1}$. Then, with $\mathbf{P}_T = \mathbf{F}\mathbf{D}_T$ we have from a straightforward calculation

$$\begin{aligned} \mathbf{Q}(\boldsymbol{\rho}_T) &= \mathbf{P}_T\mathbf{Q}(-\mathbf{c})\mathbf{D}_T^c \\ \mathbf{D}_T^c\mathbf{J}(\boldsymbol{\rho}_T)(\mathbf{D}_T^c)^{-1} &= \mathbf{I} + \mathbf{J}(-\mathbf{c})/T \end{aligned} \tag{28}$$

so that from (25) and (27)

$$\boldsymbol{\Phi}_T = \mathbf{P}_T(\mathbf{I} + \mathbf{A}/T)\mathbf{P}_T^{-1}. \tag{29}$$

Furthermore, since $z^{p-1} + \sum_{j=0}^{p-2}\theta_{T,j}z^j = \prod_{j=1}^{p-1}(z - \gamma_{T,j})$, we have $\boldsymbol{\theta}_T'\mathbf{Q}(\boldsymbol{\gamma}_T) = 0$, and similarly, $\mathbf{b}'\mathbf{Q}(-\mathbf{g}) = 0$. Now as in (28), $\mathbf{Q}(\boldsymbol{\gamma}_T) = \mathbf{P}_T\mathbf{Q}(-\mathbf{g})\mathbf{D}_T^g$ for some diagonal matrix $\mathbf{D}_T^g$ with nonzero diagonal elements, so that also $\boldsymbol{\theta}_T'\mathbf{P}_T\mathbf{Q}(-\mathbf{g}) = 0$. Since $\mathbf{Q}(-\mathbf{g})$ is of full column rank (cf. Theorem 2 of Brand (1964)), we conclude that $\boldsymbol{\theta}_T'\mathbf{P}_T$ is a scalar multiple of $\mathbf{b}'$. The last element of $\mathbf{b}'$ is equal to one, and the last element of $\boldsymbol{\theta}_T'\mathbf{P}_T$ is equal to T^{1-p}, so that

$$\boldsymbol{\theta}_T'\mathbf{P}_T = T^{1-p}\mathbf{b}'. \tag{30}$$

Finally, from $\mathbf{P}_T\mathbf{e} = T^{1-p}\mathbf{e}$,

$$\mathbf{P}_T^{-1}\mathbf{e} = T^{p-1}\mathbf{e}. \tag{31}$$

From (29), (30) and (31) it follows that the system (23) and (24) can equivalently be written as (10) and (11) with $\mathbf{Z}_{T,t} = T^{1-p}\mathbf{P}_T^{-1}\mathbf{V}_{T,t-1}$.

For an arbitrary valued matrix $\mathbf{B}$, let $||\mathbf{B}||$ its largest singular value. In the following, let T be large enough so that $|\rho_{T,j}|^2 = 1 - 2\,\mathrm{Re}(c_j)/T + |c_j|^2/T^2 \leq (1 - \frac{1}{2}\,\mathrm{Re}(c_1)/T)^2$ for all $j = 1, \ldots, p$, so that from (27), also

$$||\mathbf{I} + \mathbf{A}/T|| \leq 1 - \tfrac{1}{2}\,\mathrm{Re}(c_1)/T. \tag{32}$$

Now from (10) and (11), for any fixed integer $K > 0$,

$$T^{-1/2}(x_{T,\lceil sT\rceil} - \mu) = R_T(s) + T^{-1/2}\mathbf{b}'\sum_{t=-KT+1}^{\lceil sT\rceil}(\mathbf{I} + \mathbf{A}/T)^{\lceil sT\rceil - t}\mathbf{e}u_t$$

where $R_T(s) = T^{-1/2}\mathbf{b}'(\mathbf{I}+\mathbf{A}/T)^{\lceil sT\rceil + KT}\mathbf{Z}_{-KT}$ and $\mathbf{Z}_{-KT} = \sum_{t=0}^{\infty}(\mathbf{I}+\mathbf{A}/T)^t\mathbf{e}u_{-KT-t}$, and we write $\mathbf{Z}_t$ for $\mathbf{Z}_{T,t}$ to ease notation. Since the autocovariances of u_t are absolutely summable, the spectral density of u_t exists and is bounded on $[-\pi, \pi]$. Let a bound be $\tilde{\sigma}_u^2/(2\pi)$. For

any given T and $\mathbf{w} \in \mathbb{R}^p$, the variance of the time invariant linear filter $\mathbf{w}'\mathbf{Z}_{-KT}$ is thus weakly smaller than the variance of $\mathbf{w}'\tilde{\mathbf{Z}}_{-KT}$, where $\tilde{\mathbf{Z}}_{-KT} = \sum_{t=0}^{\infty}(\mathbf{I} + \mathbf{A}/T)^t \mathbf{e}\tilde{u}_{-KT-t}$ with $\tilde{u}_t \sim iid(0, \tilde{\sigma}_u^2)$. Furthermore

$$\mathrm{Var}[T^{-1/2}\tilde{\mathbf{Z}}_{-KT}] = \tilde{\sigma}_u^2 T^{-1} \sum_{t=0}^{\infty}(\mathbf{I} + \mathbf{A}/T)^t \mathbf{e}\mathbf{e}'(\mathbf{I} + \mathbf{A}'/T)^t$$

so that from (32)

$$||\,\mathrm{Var}[T^{-1/2}\tilde{\mathbf{Z}}_{-KT}]|| \leq \tilde{\sigma}_u^2 ||\mathbf{e}\mathbf{e}'|| T^{-1} \sum_{t=0}^{\infty}(1 - \tfrac{1}{2}\,\mathrm{Re}(c_1)/T)^{2t} = O(1).$$

Thus, $||T^{-1/2}\mathbf{Z}_{-KT}|| = O_p(1)$. Using again (32), we obtain

$$\begin{aligned} \sup_{0\leq s\leq 1} |R_T(s)| &\leq ||\mathbf{b}||\cdot||T^{-1/2}\mathbf{Z}_{-KT}|| \cdot \sup_s (1 - \tfrac{1}{2}\,\mathrm{Re}(c_1)/T)^{\lceil sT\rceil + KT} \\ &\leq ||\mathbf{b}||\cdot||T^{-1/2}\mathbf{Z}_{-KT}|| \cdot \exp[-\tfrac{1}{2}K\,\mathrm{Re}(c_1)] \end{aligned} \tag{33}$$

so that $R_T(\cdot)$ converges in probability as $K \to \infty$ in the sense that for any $\varepsilon > 0$, there exists $K = K_\varepsilon$ such that $\limsup_{T\to\infty} P(\sup_{0\leq s\leq 1} |R_T(s)| > \varepsilon) < \varepsilon$.

Furthermore, under Condition 1, $W_T(\cdot) = T^{-1/2}\sum_{t=-\lceil KT\rceil}^{\lceil \cdot T\rceil} u_t \Rightarrow W(\cdot) - W(-K)$, where W is a Wiener process on the interval $[-K, 1]$ of variance ω^2 normalized to $W(0) = 0$. By summation by parts,

$$\begin{aligned} & T^{-1/2}\mathbf{b}' \sum_{t=-\lceil KT\rceil+1}^{\lceil sT\rceil} (\mathbf{I} + \mathbf{A}/T)^{\lceil sT\rceil - t}\mathbf{e}u_t \\ = \;& \mathbf{b}'\mathbf{e}W_T(s) - \mathbf{b}'(\mathbf{I} + \mathbf{A}/T)^{\lceil sT\rceil + \lceil KT\rceil - 1}\mathbf{e}T^{-1/2}u_{-\lceil KT\rceil} \\ & + \mathbf{b}'\mathbf{A}T^{-1} \sum_{t=-\lceil KT\rceil+2}^{\lceil sT\rceil} (\mathbf{I} + \mathbf{A}/T)^{\lceil sT\rceil - t}\mathbf{e}W_T(\frac{t-1}{T}) \\ \Rightarrow \;& \mathbf{b}'\mathbf{e}(W(s) - W(-K)) + \mathbf{b}'\mathbf{A}\int_{-K}^{s} e^{\mathbf{A}(s-r)}\mathbf{e}(W(r) - W(-K))dr \\ = \;& \mathbf{b}'\int_{-K}^{s} e^{\mathbf{A}(s-r)}\mathbf{e}dW(r) = R^0(s) + J_p(s) \end{aligned}$$

where $R^0(s) = -\mathbf{b}'e^{\mathbf{A}(s+K)}\mathbf{X}(-K)$ with $\mathbf{X}(-K) \sim \mathcal{N}(0, \mathbf{\Sigma})$ independent of W as in (7), the convergence relies on the well-known identity $e^{s\mathbf{A}} = \lim_{T\to\infty}(\mathbf{I} + \mathbf{A}/T)^{\lceil sT\rceil}$ for all s, and the second equality follows from the stochastic calculus version of integration by parts. Since $\sup_{0\leq s\leq 1} ||e^{\mathbf{A}(s+K)}|| \to 0$ as $K \to \infty$, $\sup_{0\leq s\leq 1} |R^0(s)|$ converges in probability to zero as $K \to \infty$. As noted below (33), the same holds for $R_T(\cdot)$. But convergence in probability implies convergence in distribution, and K was arbitrary, so the result follows.

A.2 Proof of Theorem 2

Overview

The proof of Theorem 2 relies heavily on the framework developed by Ibragimov and Rozanov (1978), denoted IR78 in the following. As discussed there, a continuous time Gaussian process on the unit interval can be described in terms of a countably infinite sequence of random variables (cf. (42) and the discussion in the proof of Lemma 4 below), whose distribution can be expressed in terms of the spectral density of the underlying process (cf. (41) and the discussion below (42)). The challenge in the proof of Theorem 2 is to establish that the "infinite tail" of this sequence contributes negligibly to the total variation distance. Intuitively, this must hold for some appropriate definition of tail if the two measures are equivalent, and appropriate equivalence results are obtained by IR78. But the construction of this tail must be such that its contribution is negligible *uniformly* over a sufficiently rich class of potential approximating processes. To this end, the sequence of random variables (and hence its tail) is constructed as a function of the properties of two Gaussian processes whose spectral densities form an upper and lower bound on the class of potential approximating spectral density functions (cf. (37), (38) and (39)), which turns out to be suitable to obtain such a uniform bound (cf. (44), (45) and (48)). With the contribution from the tail controlled, the approximability of the distribution of the finite dimensional non-tail part of the sequence of random variables follows with some additional work from Lemmas 1 and 2 below.

We first state Lemmas 1 and 2. We write z^* for the conjugate of the complex number z, and $\mathbf{v}^*$ for the conjugate transpose of a complex vector $\mathbf{v}$.

Lemma 1 *Let $\mathcal{C}_0$ be the space of continuous real valued functions on $[0,\infty)$ which vanish at infinity. For any $\vartheta_0 \in \mathcal{C}_0$ and $\varepsilon > 0$, there exists an integer $q \geq 1$ such that $\sup_{\lambda\geq 0} |\vartheta_0(\lambda) - \vartheta(\lambda)| < \varepsilon$, where ϑ is a rational function of the form*

$$\vartheta(\lambda) = \frac{\sum_{j=0}^{q-1} e_j^n \lambda^{2j}}{\prod_{j=1}^{q}(\lambda^2 + e_j^d)} \tag{34}$$

with $e_j^d > 0$ and $e_j^n \in \mathbb{R}$, $j = 0, \ldots, q$.

Proof. Note that functions of the form ϑ form a vector subspace of $\mathcal{C}_0$ which is closed under multiplication of functions, that is, they form a sub-algebra on $\mathcal{C}_0$. It is easily seen

that this sub-algebra separates points and vanishes nowhere. The locally compact version of the Stone-Weierstrass Theorem thus implies the result. ■

Lemma 2 *Let $\xi_p(\lambda^2)$ be a polynomial with real coefficients of order $p-1$ in λ^2 such that $\xi_p(\lambda^2) > 0$ for all $\lambda \in \mathbb{R}$, and with unit coefficient on $(\lambda^2)^{p-1}$. Then there exists polynomial b of order $p-1$ of the form $b(z) = \prod_{j=1}^{p-1}(z+g_j)$ with g_j as described in Condition 1 (ii) such that $\xi_p(\lambda^2) = |b(i\lambda)|^2$ for all $\lambda \in \mathbb{R}$.*

Proof. By the fundamental theorem of algebra, and since $\xi_p(\lambda^2) > 0$ for all $\lambda \in \mathbb{R}$, $\xi_p(\lambda^2) = \prod_{j=1}^{p-1}(\lambda^2+\eta_j)$, where the η_j's are of two types: real and positive, or complex with positive real part, and in conjugate pairs. Now for $0 < \eta_j \in \mathbb{R}$, $\lambda^2+\eta_j = |i\lambda+g_j|^2$ with $g_j = \sqrt{\eta_j}$. For $\eta_j = \eta_{j'}^* \in \mathbb{C}$ for $j \neq j'$,

$$(\lambda^2+\eta_j)(\lambda^2+\eta_{j'}) = \lambda^4 + 2\,\mathrm{Re}(\eta_j)\lambda^2 + |\eta_j|^2 = |i\lambda+g_j|^2|i\lambda+g_j^*|^2$$

where $\sqrt{2}g_j = \sqrt{|\eta_j| + \mathrm{Re}(\eta_j)} + \sqrt{|\eta_j| - \mathrm{Re}(\eta_j)}i$. ■

Without loss of generality, assume $\omega^2 = 2\pi$. In the following, we write G_1 for G, and f_1 for its spectral density. Let $f_0(\lambda) = (1+\lambda^2)^{-1}$ be the spectral density of the Ornstein-Uhlenbeck process with mean reversion parameter equal to unity, denoted G_0, and let

$$\delta_0 = \tfrac{1}{2}\min(\inf_\lambda(1+\lambda^2)f_1(\lambda), 1). \tag{35}$$

Let P_0 and P_1 be the measures of G_0 and G_1, respectively. By Theorem III.17 of IR78, equivalence of P_0 and P_1 implies

$$\int (\frac{f_1(\lambda)}{f_0(\lambda)} - 1)^2 d\lambda < \infty \tag{36}$$

and here and below, integrals are over the entire real line unless indicated otherwise.

Define

$$\overline{f}(\lambda) = \max(f_1(\lambda), f_0(\lambda)) + \frac{\delta_0}{(1+\lambda^2)^2} \tag{37}$$

$$\underline{f}(\lambda) = \min(f_1(\lambda), f_0(\lambda)) - \frac{\delta_0}{(1+\lambda^2)^2} \tag{38}$$

and let f_2 be some function satisfying

$$\underline{f}(\lambda) \leq f_2(\lambda) \leq \overline{f}(\lambda) \text{ for all } \lambda. \tag{39}$$

From (36),

$$\int (\frac{\underline{f}(\lambda)}{f_0(\lambda)} - 1)^2 d\lambda < \infty, \ \int (\frac{\overline{f}(\lambda)}{f_0(\lambda)} - 1)^2 d\lambda < \infty \tag{40}$$

so that also for all f_2 satisfying (39), $\int (f_2(\lambda)/f_0(\lambda) - 1)^2 d\lambda < \infty$. Since $\overline{f}$, $\underline{f}$ and f_2 are nonnegative integrable real functions, there exist corresponding correlation functions that are positive definite. By the development in Section I.2 of IR78, there hence exist corresponding stationary Gaussian processes $\overline{G}$, $\underline{G}$ and G_2 with spectral densities $\overline{f}$, $\underline{f}$ and f_2 and measures $\overline{P}$, $\underline{P}$ and P_2, respectively. Theorem III.17 of IR78 and (40) implies that $\overline{P}$, $\underline{P}$ and P_2 are equivalent to P_0, and hence also to P_1.

For $\psi, \varphi : \mathbb{R} \mapsto \mathbb{C}$ functions of the type $\psi(\lambda) = \sum_{l=1}^{k} c_l e^{i\lambda t_l}$ for some $k \geq 1$, $t_l \in [0, 1]$ and $c_l \in \mathbb{R}$, define the inner product

$$\langle \psi, \varphi \rangle_{F_1} = \int \psi(\lambda) \varphi(\lambda)^* f_1(\lambda) d\lambda. \tag{41}$$

Let $L(F_1)$ be the corresponding Hilbert space. Analogously, define the inner products $\langle \psi, \varphi \rangle_{F_2}$, $\langle \psi, \varphi \rangle_{\overline{F}}$ and $\langle \psi, \varphi \rangle_{\underline{F}}$, and corresponding Hilbert spaces $L(F_2)$, $L(\overline{F})$ and $L(\underline{F})$. Since the measures P_1, P_2, $\overline{P}$ and $\underline{P}$ are equivalent, so are the Hilbert spaces, as noted on page 71 of IR78. Define the linear operator $A : L(F_1) \mapsto L(F_2)$ via $A\psi = \psi$, let A^* be its adjoint, and define the self-adjoint operator $\Delta : L(F_1) \mapsto L(F_1)$ via $\Delta\psi = \psi - A^* A\psi$, so that

$$\langle \Delta\psi, \varphi \rangle_{F_1} = \langle \psi, \varphi \rangle_{F_1} - \langle \psi, \varphi \rangle_{F_2}$$

and analogously for $\overline{\Delta}$ and $\underline{\Delta}$ (that is, $\langle \overline{\Delta}\psi, \varphi \rangle_{F_1} = \langle \psi, \varphi \rangle_{F_1} - \langle \psi, \varphi \rangle_{\overline{F}}$ and $\langle \underline{\Delta}\psi, \varphi \rangle_{F_1} = \langle \psi, \varphi \rangle_{F_1} - \langle \psi, \varphi \rangle_{\underline{F}}$). By Theorem III.4 of IR78, equivalence of the measures P_1, P_2, $\overline{P}$ and $\underline{P}$ implies that the operators Δ, $\overline{\Delta}$ and $\underline{\Delta}$ are Hilbert-Schmidt.

Let ψ_k be an arbitrary orthonormal sequence in $L(F_1)$, and define the $n \times 1$ vector $\boldsymbol{\eta}_n$ of Gaussian complex valued random variables

$$\eta(\psi_k) = \int \psi_k(\lambda) d\Phi_l(\lambda) \text{ for } k = 1, \ldots n \tag{42}$$

where Φ_l is the stochastic spectral measure such that $G_l(s) = \int e^{i\lambda s} d\Phi_l(\lambda)$, $l = 1, 2$ (cf. Chapter I.6 of IR78). Then $E[\eta(\psi_k)] = 0$ under both P_1 and P_2, $E[\eta(\psi_j)\eta(\psi_k)] = \langle \psi_j, \psi_k \rangle_{F_1} = \mathbf{1}[j = k]$ under P_1, and $E[\eta(\psi_j)\eta(\psi_k)] = \langle \psi_j, \psi_k \rangle_{F_2}$ under P_2. Thus, under P_1, $\boldsymbol{\eta}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and under P_2, $\boldsymbol{\eta}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_n)$, where $\boldsymbol{\Sigma}_n$ has elements $\langle \psi_j, \psi_k \rangle_{F_2}$. Since P_1 and P_2 are equivalent, $\boldsymbol{\Sigma}_n$ is positive definite for any n (cf. page 76 of IR78). Let $\mathbf{v}_{kn} \in \mathbb{C}^n$, $k = 1, \ldots, n$

be a set of eigenvectors of $\mathbf{\Sigma}_n$ with associated eigenvalues σ_{kn}^2, so that $\mathbf{v}_{kn}^* \boldsymbol{\eta}_n \sim iid\mathcal{N}(0,1)$ under P_1, and $\mathbf{v}_{kn}^* \boldsymbol{\eta}_n$ are independent $\mathcal{N}(0, \sigma_{kn}^2)$ under P_2. Let d_n be the entropy distance between the distribution of $\boldsymbol{\eta}_n$ under P_1 and P_2, that is the sum of the two corresponding Kullback-Leibler divergences. By a straightforward calculation (cf. equation (III.2.4) of IR78), $d_n = \frac{1}{2}\sum_{k=1}^n [(1/\sigma_{kn}^2 - 1) + (\sigma_{kn}^2 - 1)]$. Define

$$D_n = \sum_{k=1}^n (1 - \sigma_{kn}^2)^2$$

and with $\lambda_k(\mathbf{B})$ denoting the kth largest eigenvalue of the Hermitian matrix $\mathbf{B}$, we have

$$D_n = \sum_{k=1}^n (1 - \lambda_k(\mathbf{\Sigma}_n))^2 = \sum_{k=1}^n \lambda_k((\mathbf{I}_n - \mathbf{\Sigma}_n)^2) = \mathrm{tr}((\mathbf{I}_n - \mathbf{\Sigma}_n)^2)$$

so that

$$D_n = \sum_{j,k=1}^n |\langle \psi_j, \psi_k \rangle_{F_1} - \langle \psi_j, \psi_k \rangle_{F_2}|^2 = \sum_{j,k=1}^n |\langle \Delta\psi_j, \psi_k \rangle_{F_1}|^2. \tag{43}$$

The following straightforward Lemma establishes a useful relationship between d_n and D_n.

Lemma 3 *For any $0 < \delta < 1/4$, $D_n < \delta$ implies $d_n < \delta$.*

Proof. Note that $\sum_{k=1}^n (\sigma_k^2 - 1)^2 < 1/4$ implies $1/2 < \sigma_k^2 < 3/2$ for all $k = 1, \ldots, n$, but for such σ_k^2, $\frac{1}{2}[(\frac{1}{\sigma_k^2} - 1) + (\sigma_k^2 - 1)] \leq (\sigma_k^2 - 1)^2$, which implies the result. ■

Let $\overline{\mathbf{\Sigma}}_n$ be the $n \times n$ Hermitian matrix with elements $\langle \psi_j, \psi_k \rangle_{\overline{F}}$. Then for any $\mathbf{v} = (v_1, \ldots, v_n)' \in \mathbb{C}^n$, $\mathbf{v}^*(\overline{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_n)\mathbf{v} = \sum_{j,k=1}^n v_k^* v_j (\langle \psi_j, \psi_k \rangle_{\bar{F}} - \langle \psi_j, \psi_k \rangle_{F_2}) = \int ||\sum_{j=1}^n v_j \psi_j||^2 (\bar{f}(\lambda) - f_2(\lambda)) d\lambda \geq 0$ from (39). Therefore, by Weyl's inequality, $\overline{\sigma}_{kn}^2 = \lambda_k(\overline{\mathbf{\Sigma}}_n) \geq \lambda_k(\mathbf{\Sigma}_n) + \lambda_n(\overline{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_n) \geq \lambda_k(\mathbf{\Sigma}_n) = \sigma_{kn}^2$ for all k, so that

$$\overline{D}_n = \sum_{j,k=1}^n |\langle \overline{\Delta}\psi_j, \psi_k \rangle_{F_1}|^2 = \sum_{k=1}^n (1 - \overline{\sigma}_{kn}^2)^2 \geq \sum_{k=1}^n \mathbf{1}[\sigma_{kn}^2 > 1](1 - \sigma_{kn}^2)^2.$$

By an analogous argument, also

$$\underline{D}_n = \sum_{j,k=1}^n |\langle \underline{\Delta}\psi_j, \psi_k \rangle_{F_1}|^2 \geq \sum_{k=1}^n \mathbf{1}[\sigma_{kn}^2 < 1](1 - \sigma_{kn}^2)^2$$

so that

$$D_n \leq \overline{D}_n + \underline{D}_n. \tag{44}$$

Now let $\overline{\varphi}_k$ be a complete set of eigenvectors of the operator $\overline{\Delta}$, with associated eigenvalues $1-\overline{\sigma}_k^2$, that is $\overline{\Delta}\overline{\varphi}_k = (1-\overline{\sigma}_k^2)\overline{\varphi}_k$, and $\overline{\varphi}_k$ form an orthonormal basis in $L(F_1)$. Define $\underline{\varphi}_k$ and $\underline{\sigma}_k^2$ analogously relative to the operator $\underline{\Delta}$. Since $\overline{\Delta}$ and $\underline{\Delta}$ are Hilbert-Schmidt, $\sum_{k=1}^{\infty}(1-\overline{\sigma}_k^2)^2 < \infty$ and $\sum_{k=1}^{\infty}(1-\underline{\sigma}_k^2)^2 < \infty$, so that for any $\epsilon > 0$, there exists n_ϵ such that $\sum_{k=n_\epsilon}^{\infty}(1-\overline{\sigma}_k^2)^2 < \epsilon/2$ and $\sum_{k=n_\epsilon}^{\infty}(1-\underline{\sigma}_k^2)^2 < \epsilon/2$. Let $\overline{L}_\epsilon^0 \subset L(F_1)$ and $\underline{L}_\epsilon^0 \subset L(F_1)$ be the spaces spanned by $\overline{\varphi}_k$ and $\underline{\varphi}_k$, $k = 1, \ldots, n_\epsilon$, respectively, and let L_ϵ^1 be the orthogonal complement of $L_\epsilon^0 = \overline{L}_\epsilon^0 \cup \underline{L}_\epsilon^0$ relative to $\langle \cdot, \cdot \rangle_{F_1}$, so that $L(F_1) = L_\epsilon^0 \cup L_\epsilon^1$. Note that L_ϵ^0 and L_ϵ^1 do not depend on f_2. Let $\overline{L}_\epsilon^1$ be the space spanned by $\overline{\varphi}_k$, $k = n_\epsilon + 1, n_\epsilon + 2, \ldots$ For any orthonormal sequence ψ_k in L_ϵ^1, since $L_\epsilon^1 \subset \overline{L}_\epsilon^1$

$$\overline{D}_n = \sum_{j,k=1}^{n} |\langle \overline{\Delta}\psi_j, \psi_k \rangle_{F_1}|^2 \leq \sum_{j=1}^{n} ||\overline{\Delta}\psi_j||_{\overline{L}_\epsilon^1}^2 = \sum_{j=1}^{n} \sum_{k=n_\epsilon+1}^{\infty} |\langle \overline{\Delta}\psi_j, \overline{\varphi}_k \rangle_{F_1}|^2 = \sum_{k=n_\epsilon+1}^{\infty} \sum_{j=1}^{n} |\langle \overline{\Delta}\overline{\varphi}_k, \psi_j \rangle_{F_1}|^2$$

where the inequality follows from Bessel's inequality. A further application yields

$$\sum_{k=n_\epsilon+1}^{\infty} \sum_{j=1}^{n} |\langle \overline{\Delta}\overline{\varphi}_k, \psi_j \rangle_{F_1}|^2 \leq \sum_{k=n_\epsilon+1}^{\infty} ||\overline{\Delta}\overline{\varphi}_k||_{\overline{L}_\epsilon^1}^2 = \sum_{j,k=n_\epsilon+1}^{\infty} (1-\overline{\sigma}_k^2)^2 |\langle \overline{\varphi}_j, \overline{\varphi}_k \rangle_{F_1}|^2 = \sum_{k=n_\epsilon+1}^{\infty} (1-\overline{\sigma}_k^2)^2$$

with the right-hand side bounded above by $\epsilon/2$ by the definition of n_ϵ. Thus $\overline{D}_n < \epsilon/2$ and, by the analogous argument, also $\underline{D}_n \leq \epsilon/2$. Thus, from (44), for any orthonormal sequence ψ_k in L_ϵ^1,

$$D_n \leq \epsilon. \tag{45}$$

Now let ψ_k^ϵ, $k = 1, \ldots, m_\epsilon \leq 2n_\epsilon$ be an orthonormal basis of L_ϵ^0, and let ψ_k^ϵ, $k = m_\epsilon + 1, m_\epsilon + 2, \ldots$ be an orthonormal basis of L_ϵ^1, so that ψ_k^ϵ, $k = 1, 2, \ldots$ is an orthonormal basis of $L(F_1)$. Note that the sequence ψ_k^ϵ does not depend on f_2. Let $\mathfrak{U}_m^\epsilon$ be the σ-field generated by the Gaussian random variables $\eta(\psi_k^\epsilon)$ as defined in (42) for $k = 1, \ldots, m$, $l = 1, 2$, and let $\mathfrak{U}^\epsilon$ be the σ-field generated by $\eta(\psi_k^\epsilon)$, $k = 1, 2, \ldots$. Define

$$D_m^\epsilon = \sum_{j,k=1}^{m} |\langle \Delta\psi_j^\epsilon, \psi_k^\epsilon \rangle_{F_1}|^2.$$

We have the following Lemma.

Lemma 4 *For all $0 < \epsilon_0 < 1/2$, $\sup_m D_m^\epsilon < \epsilon_0^2$ implies that the total variation distance between P_1 and P_2 is smaller than ϵ_0.*

Proof. As discussed on page 65 of IR78, the distribution on the σ-field $\mathfrak{U}^\epsilon$ equivalently characterizes the distribution of G_l relative to the σ-fields generated by the cylindric sets of the paths $G_l(\cdot)$ under P_l, $l = 1, 2$, so it suffices to show that

$$\sup_{\mathcal{A}\in\mathfrak{U}^\epsilon} |P_2(\mathcal{A}) - P_1(\mathcal{A})| \leq \epsilon_0. \tag{46}$$

Let d_m^ϵ be the entropy distance between the distribution of $\eta(\psi_k^\epsilon)$, $k = 1, \ldots, m$ under P_1 and P_2. By Lemma 3, $d_m^\epsilon \leq \epsilon_0^2$. Thus, by Pinsker's inequality

$$\sup_{\mathcal{A}_m\in\mathfrak{U}_m^\epsilon} |P_2(\mathcal{A}_m) - P_1(\mathcal{A}_m)| \leq \epsilon_0 \text{ for all } m. \tag{47}$$

Now suppose (46) does not hold. Then there exists $\mathcal{A} \in \mathfrak{U}^\epsilon$ such that $P_2(\mathcal{A}) - P_1(\mathcal{A}) > \epsilon_0$. Construct a sequence of events $\mathcal{A}_m \in \mathfrak{U}_m^\epsilon$ such that $P_l(\mathcal{A}_m \ominus \mathcal{A}) \to 0$ for $l = 1, 2$ as on page 77 of IR78, where $\mathcal{A}_m \ominus \mathcal{A}$ is the symmetric difference $\mathcal{A}_m \ominus \mathcal{A} = (\mathcal{A}_m \cup \mathcal{A})\backslash(\mathcal{A}_m \cap \mathcal{A})$. Then from $\mathcal{A} \subseteq \mathcal{A}_m \cup (\mathcal{A}_m \ominus \mathcal{A})$ and $\mathcal{A}_m \subseteq \mathcal{A} \cup (\mathcal{A}_m \ominus \mathcal{A})$, we have $|P_l(\mathcal{A}) - P_l(\mathcal{A}_m)| \leq P_l(\mathcal{A}_m \ominus \mathcal{A})$ for $l = 1, 2$. We thus obtain $P_2(\mathcal{A}_m) - P_1(\mathcal{A}_m) \to P_2(\mathcal{A}) - P_1(\mathcal{A}) > \epsilon_0$, contradicting (47), and the lemma is proved. ■

Given that the choice of $0 < \epsilon$ was arbitrary, in light of Lemma 4 it suffices to show that for some CARMA implied f_2 satisfying (39), $\sup_m D_m^\epsilon < 2\epsilon$, say. Now for all $m > m_\epsilon$, from (43)

$$\begin{aligned} D_m^\epsilon &\leq \sum_{j,k=m_\epsilon+1}^{m} |\langle \Delta\psi_j^\epsilon, \psi_k^\epsilon\rangle_{F_1}|^2 + 2\sum_{j=1}^{m_\epsilon}\sum_{k=1}^{m} |\langle \Delta\psi_j^\epsilon, \psi_k^\epsilon\rangle_{F_1}|^2 \\ &\leq \epsilon + 2\sum_{j=1}^{m_\epsilon}\sum_{k=1}^{m} |\langle \Delta\psi_j^\epsilon, \psi_k^\epsilon\rangle_{F_1}|^2 \end{aligned} \tag{48}$$

where the second inequality follows from (45). Further

$$\begin{aligned} \sum_{k=1}^{m} |\langle \Delta\psi_j^\epsilon, \psi_k^\epsilon\rangle_{F_1}|^2 &= \sum_{k=1}^{m} |\langle (\frac{f_2}{f_1} - 1)\psi_j^\epsilon, \psi_k^\epsilon\rangle_{F_1}|^2 \\ &\leq \langle (\frac{f_2}{f_1} - 1)\psi_j^\epsilon, (\frac{f_2}{f_1} - 1)\psi_j^\epsilon\rangle_{F_1}^2 \\ &= \int (\frac{f_2(\lambda)}{f_1(\lambda)} - 1)^2 |\psi_j^\epsilon(\lambda)|^2 f_1(\lambda) d\lambda \end{aligned}$$

where the inequality follows from Bessel's inequality by viewing $L(F_1)$ as a subspace of the Hilbert space of square integrable functions with inner-product $\langle\cdot,\cdot\rangle_{F_1}$. Thus

$$\sup_m D_m^\epsilon \leq \epsilon + 2\sum_{j=1}^{m_\epsilon} \int (\frac{f_2(\lambda)}{f_1(\lambda)} - 1)^2 |\psi_j^\epsilon(\lambda)|^2 f_1(\lambda) d\lambda. \tag{49}$$

From equation (II.1.3) of IR78 and by condition (ii) of Theorem 2, every $\psi \in L(F_1)$ can be represented in the form $\psi(\lambda) = c_0 + (1+i\lambda)\int_0^1 e^{i\lambda t}c(t)dt$ for some real c_0 and some square integrable function $c : [0,1] \mapsto \mathbb{R}$. Thus $|\psi(\lambda)| \leq |c_0| + \sqrt{1+\lambda^2}\int_0^1 |c(t)|dt$, so that $\sup_\lambda |\psi(\lambda)|^2 f_1(\lambda) < \infty$. Thus, (49) implies that for some $M_\epsilon < \infty$ that does not depend on f_2, for all f_2 satisfying (39),

$$\sup_m D_m^\epsilon \leq \epsilon + M_\epsilon \int (\frac{f_2(\lambda)}{f_1(\lambda)} - 1)^2 d\lambda. \tag{50}$$

It thus suffices to show that there exists a CARMA implied f_2 satisfying (39) that makes $\int_0^\infty (f_2(\lambda)/f_1(\lambda) - 1)^2 d\lambda$ arbitrarily small.

Let $h_1(\lambda) = f_1(\lambda)/f_0(\lambda) - 1$ and $h_2(\lambda) = f_2(\lambda)/f_0(\lambda) - 1$. Recalling the definition of δ_0 in (35), we have

$$\int_0^\infty (\frac{f_2(\lambda)}{f_1(\lambda)} - 1)^2 d\lambda \leq \tfrac{1}{4}\delta_0^{-2} \int_0^\infty \left(h_1(\lambda) - h_2(\lambda)\right)^2 d\lambda$$

and it suffices to show that for any $\epsilon_1 > 0$, there exists a CARMA implied h_2 such that $\int_0^\infty \left(h_1(\lambda) - h_2(\lambda)\right)^2 d\lambda < 2\epsilon_1$.

For any $\tilde{h}_1(\lambda)$, from $(a-b)^2 \leq 2(a^2+b^2)$,

$$\begin{aligned}\int_0^\infty \left(h_1(\lambda) - h_2(\lambda)\right)^2 d\lambda &= \int_0^\infty \Big(h_1(\lambda) - \tilde{h}_1(\lambda) - h_2(\lambda) + \tilde{h}_1(\lambda)\Big)^2 d\lambda \\ &\leq 2\int_0^\infty (h_1(\lambda) - \tilde{h}_1(\lambda))^2 d\lambda + 2\int_0^\infty (h_2(\lambda) - \tilde{h}_1(\lambda))^2 d\lambda.\end{aligned}$$

By (36), $\int_0^\infty h_1(\lambda)^2 d\lambda < \infty$. Thus, there exists $K < \infty$ such that $\int_K^\infty h_1(\lambda)^2 d\lambda < \epsilon_1/2$. Let $\chi_K(\lambda) = 1$ for $\lambda \leq K$, $\chi_K(\lambda) = 0$ for $\lambda \geq K+1$ and $\chi_K(\lambda) = K + 1 - \lambda$ otherwise, and define $\tilde{h}_1(\lambda) = \chi_K(\lambda)h_1(\lambda)$. Then $\int_0^\infty (h_1(\lambda) - \tilde{h}_1(\lambda))^2 d\lambda \leq \epsilon_1/2$, and since f_1 is continuous by standard Fourier arguments (see, for instance, Proposition 4.1 on page 87 in Stein and Shakarchi (2005)), so is $\tilde{h}_1$. It thus suffices to show that there exists a CARMA implied h_2 that makes $\int_0^\infty (h_2(\lambda) - \tilde{h}_1(\lambda))^2 d\lambda$ smaller than $\epsilon_1/2$.

Now

$$\int_0^\infty (h_2(\lambda) - \tilde{h}_1(\lambda))^2 d\lambda = \int_0^\infty (1+\lambda^2)^{-2} (\vartheta_2(\lambda) - \tilde{\vartheta}_1(\lambda))^2 d\lambda$$

with $\vartheta_2(\lambda) = (1+\lambda^2)h_2(\lambda)$ and $\tilde{\vartheta}_1(\lambda) = (1+\lambda^2)\tilde{h}_1(\lambda)$. Note that $\tilde{\vartheta}_1$ is continuous, and $\lim_{\lambda\to\infty} \tilde{\vartheta}_1(\lambda) = 0$. Thus, by Lemma 1, for any $\delta > 0$, there exists an integer q and a rational function ϑ_2 of the form (34) such that

$$\sup_\lambda |\vartheta_2(\lambda) - \tilde{\vartheta}_1(\lambda)| < \delta. \tag{51}$$

We have $\int_0^\infty (1+\lambda^2)^{-2}(\vartheta_2(\lambda)-\tilde{\vartheta}_1(\lambda))^2 d\lambda \le \delta^2 \int_0^\infty (1+\lambda^2)^{-2} d\lambda$, which can be made arbitrarily small by choosing δ small. From the definitions of ϑ_2 and h_2, we have

$$f_2(\lambda) = \frac{1}{1+\lambda^2} + \frac{\vartheta_2(\lambda)}{(1+\lambda^2)^2}$$

so the implied $f_2(\lambda)$ is a rational function in λ^2 of degree $p = q+2$ in the denominator and $p-1$ in the numerator.

Furthermore, for all $\delta < \delta_0$, we have uniformly in λ,

$$\begin{aligned} f_2(\lambda) &\le \frac{1}{1+\lambda^2} + \frac{\tilde{\vartheta}_1(\lambda) + \delta_0}{(1+\lambda^2)^2} \\ &= \frac{1}{1+\lambda^2} + \frac{\chi_K(\lambda)((1+\lambda^2)^2 f_1(\lambda) - (1+\lambda^2)) + \delta_0}{(1+\lambda^2)^2} \\ &= \chi_K(\lambda) f_1(\lambda) + (1-\chi_K(\lambda)) f_0(\lambda) + \frac{\delta_0}{(1+\lambda^2)^2} \le \overline{f}(\lambda) \end{aligned}$$

and similarly,

$$\begin{aligned} f_2(\lambda) &\ge \frac{1}{1+\lambda^2} + \frac{\tilde{\vartheta}_1(\lambda) - \delta_0}{(1+\lambda^2)^2} \\ &= \chi_K(\lambda) f_1(\lambda) + (1-\chi_K(\lambda)) f_0(\lambda) - \frac{\delta_0}{(1+\lambda^2)^2} \ge \underline{f}(\lambda) \end{aligned}$$

so that f_2 satisfies (39). In particular, since $\underline{f}(\lambda) > 0$ for all λ, the numerator of $f_2(\lambda)$ is a positive rational function, so by Lemma 2, f_2 has the form of the spectral density of a CARMA$(p, p-1)$ process.

A.3 Derivation of (22)

From (9), we obtain

$$\left(\frac{\partial^2 \ln f_p(\lambda)}{\partial \lambda^2}\right)^2 = \left(2\sum_{j=1}^{p-1} \frac{g_j^2 - \lambda^2}{(g_j^2+\lambda^2)^2} - 2\sum_{j=1}^{p} \frac{c_j^2 - \lambda^2}{(c_j^2+\lambda^2)^2}\right)^2. \tag{52}$$

By a direct calculation, for any $c, g \in \mathbb{C}$ with positive real part

$$\int_{-\infty}^{\infty} \frac{g^2-\lambda^2}{(g^2+\lambda^2)^2} \frac{c^2-\lambda^2}{(c^2+\lambda^2)^2} d\lambda = \frac{2\pi}{(c+g)^3}. \tag{53}$$

Equation (22) now follows from expanding (52) and applying (53).

A.4 Computational Details

A.4.1 Limited-Information Marginal Likelihood under Improper Prior for (μ, ω)

Let $\omega^2 \mathbf{\Sigma}_N$ be the $N \times N$ covariance matrix of $\{J_N(j/N)\}_{j=1}^N$. Standard arguments show that with an (improper) prior on (μ, ω) proportional to $1/\omega$, the marginal likelihood of $\mathbf{x}_T = (x_{T,\lceil T/N \rceil}, x_{T,\lceil 2T/N \rceil}, \ldots x_{T,T})'$ under the approximation (12) of Corollary 1 is given by

$$C(\boldsymbol{\iota}' \mathbf{\Sigma}_N^{-1} \boldsymbol{\iota} \det \mathbf{\Sigma}_N)^{-1/2} (\mathbf{x}_T' \mathbf{\Sigma}_N^{-1} \mathbf{x}_T - \mathbf{x}_T' \mathbf{\Sigma}_N^{-1} \boldsymbol{\iota} (\boldsymbol{\iota}' \mathbf{\Sigma}_N^{-1} \boldsymbol{\iota})^{-1} \boldsymbol{\iota}' \mathbf{\Sigma}_N^{-1} \mathbf{x}_T)^{-(N-1)/2} \tag{54}$$

where C does not depend on $\mathbf{\Sigma}_N$ and $\boldsymbol{\iota}$ is a $N \times 1$ vector of ones.

Under approximation (16), the terms in this expression may be obtained from the Kalman filter described in Section 4: As usual, $\mathbf{x}_T' \mathbf{\Sigma}_N^{-1} \mathbf{x}_T$ is the sum of squared prediction errors for the N non-missing observations $x^0_{T_0,\lceil jT_0/N \rceil} = x_{T,\lceil jT/N \rceil}$, $j = 1, \ldots, N$, normalized by their conditional variances, and $\det \mathbf{\Sigma}_N$ is the product of these conditional variances. Note that the conditional variances do not depend on the value of the observations. The quadratic form $\boldsymbol{\iota}' \mathbf{\Sigma}_N^{-1} \boldsymbol{\iota}$ is thus recognized as sum of squared prediction errors for N non-missing "dummy observations" $x^0_{T_0,\lceil jT_0/N \rceil} = 1$, $j = 1, \ldots, N$, and $\boldsymbol{\iota}' \mathbf{\Sigma}_N^{-1} \mathbf{x}_T$ is the sum of the product of these two prediction errors, both normalized by the conditional variances. The terms required to evaluate (54) conditional on $\mathbf{h}$ may thus conveniently be obtained from a single Kalman sweep with two states corresponding to the two sets of observations $x^0_{T_0,\lceil jT_0/N \rceil} = x_{T,\lceil jT/N \rceil}$ and $x^0_{T_0,\lceil jT_0/N \rceil} = 1$, $j = 1, \ldots, N$.

A.4.2 Determination of π_τ

Let $\pi_\tau(\tau) = \tau^2 \tilde{\pi}_\tau(\tau)$. The prior density of $\tau(\mathbf{h})$ implied by the prior $\pi_b(\mathbf{h}) \tau(\mathbf{h})^2 \mathbf{1}[3 \leq \tau(\mathbf{h}) \leq 50]$ on $\mathbf{h} \in [0, N\pi]^{2p-1}$ is then inversely proportional to the desired additional component $\tilde{\pi}_\tau$. We estimate $\tilde{\pi}_\tau$ from 200,000 draws from the prior $\pi_b(\mathbf{h}) \tau(\mathbf{h})^2 \mathbf{1}[3 \leq \tau(\mathbf{h}) \leq 50]$ with a step function on $[3, 50]$ with 40 equal sized steps.

A.4.3 Posterior Simulation

We employ a random walk Metropolis-Hastings algorithm to obtain a Markov Chain of draws from the posterior for $\mathbf{h}$, and thus $\tau(\mathbf{h})$. The proposed moves are Gaussian with identity covariance scaled to induce an acceptance rate of approximately 30%. To ensure the numerical stability of the state space system (14)-(15), we avoid values of $\mathbf{h}$ that lead to

$\max_{1\leq k,j\leq p}|(1-c_j/T_0)(1-c_k/T_0)|>0.999$. This is mostly binding for complex values of c_j, c_k with very small real part and very large imaginary part.

Reported results are based on the combined output from 20 independent chains with 100,000 draws each. Total computing time for a given p is about one minute on a modern workstation in a Fortran implementation.

The Bayes factors between model p and $p-1$ for $p=2,\ldots,5$ are estimated via Bridge sampling as suggested by Meng and Wong (1996) using the posterior draws from each model p, $p=1,\ldots,5$. These factors are then multiplied to obtain the Bayes factors relative to the LTU model with $p=1$.

A.5 Robustness of the Empirical Results in Section 5.2

In this Section we investigate the sensitivity of the empirical results reported in the main text to the smoothness parameter κ, and to alternative definitions of the half-life.

Recall that we set $\kappa=200$ for the coefficient that penalizes large average values of the squared second derivative of the log-spectrum. Here we consider $\kappa=500$ (imposing more smoothness) and $\kappa=0$ (imposing no smoothness).

We also consider alternative measurements for the persistence of the real exchange rates. The half-life measurement is a scalar summary of the persistence properties of J_p, as embodied in the correlation function $\gamma_p(r)/\gamma_p(0)$, $r\geq\infty$, or, equivalently, in the normalized spectrum $f_{J_p}(\lambda)/\omega^2$, and as such is necessarily imperfect.

The first two alternative definitions are simple generalizations of the half-life measure (20) as a fraction of the sample size to

$$\inf_r\left\{r:\left|\frac{\mathbf{b}'e^{\mathbf{A}|s|}\mathbf{\Sigma}\mathbf{b}}{\mathbf{b}'\mathbf{\Sigma}\mathbf{b}}\right|\leq\eta\text{ for all }s\geq r\right\}$$

for $\eta\neq 1/2$. In particular, we compute results for $\eta=1/4$ and $\eta=\sqrt{1/2}$, the "quarter-life" and "$\sqrt{1/2}$-life" analogues of the definition in the main text. To make the results directly comparable to those reported in the main text, we multiply these by $1/2$ and 2, respectively, so that in a LTU ($p=1$) specification, these definitions all coincide, at least as $T\to\infty$. We numerically approximate these using the same device as in (21).

We also consider two alternative persistence measurements that are continuous functions of the GLTU parameters. The first is based on $d=a(0)/b(0)=\prod_{j=1}^{p}c_j/\prod_{j=1}^{p-1}g_j$. As discussed in the main text, d characterizes the limit of the sum of the AR(∞) coefficients.

Table 3: Bayes Factors relative to $p = 1$ under Variants of Baseline Prior

	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$
more smoothness, $\kappa = 500$	1.0	23.1	18.0	20.9	17.3
no smoothness, $\kappa = 0$	1.0	74.7	49.9	68.6	51.1
half of quarter-life	1.0	69.6	45.1	56.8	40.3
double of $\sqrt{1/2}$-life	1.0	13.8	10.7	13.2	10.9
sum of AR(∞)	1.0	10.1	11.5	14.7	12.7
variance ratio R	1.0	67.8	43.3	57.1	40.2

From (9), $d^2 = \omega^2/(2\pi f_{J_p}(0)) = \omega^2/\int_{-\infty}^{\infty}\gamma_p(r)dr$, so d is also recognized as the square root of the ratio of the *innovation* variance and the "long-run variance" of J_p. As a final persistence measure we consider the ratio of the *unconditional* variance and the long-run variance of J_p,

$$R = \frac{\gamma_p(0)}{\int_{-\infty}^{\infty}\gamma_p(r)dr} = d^2\mathbf{b}'\mathbf{\Sigma}\mathbf{b}/\omega^2 \approx d^2\mathbf{b}'\mathbf{\Omega}^0_{T_0}\mathbf{b}/(\omega^2 T_0).$$

Note that for $p = 1$, $\mathbf{b} = 1$ and $\mathbf{\Sigma}/\omega^2 = 1/(2c_1)$, so $R = c_1/2$, and recall that the half-life for $p = 1$ as a fraction of the sample size is equal to $(\ln 2)/c_1$. To make these two additional measures more directly comparable to the baseline half-life measure, we monotonically transform them via

$$\tau_d = \frac{\ln 2}{d}, \quad \tau_R = \frac{\ln 2}{2R}$$

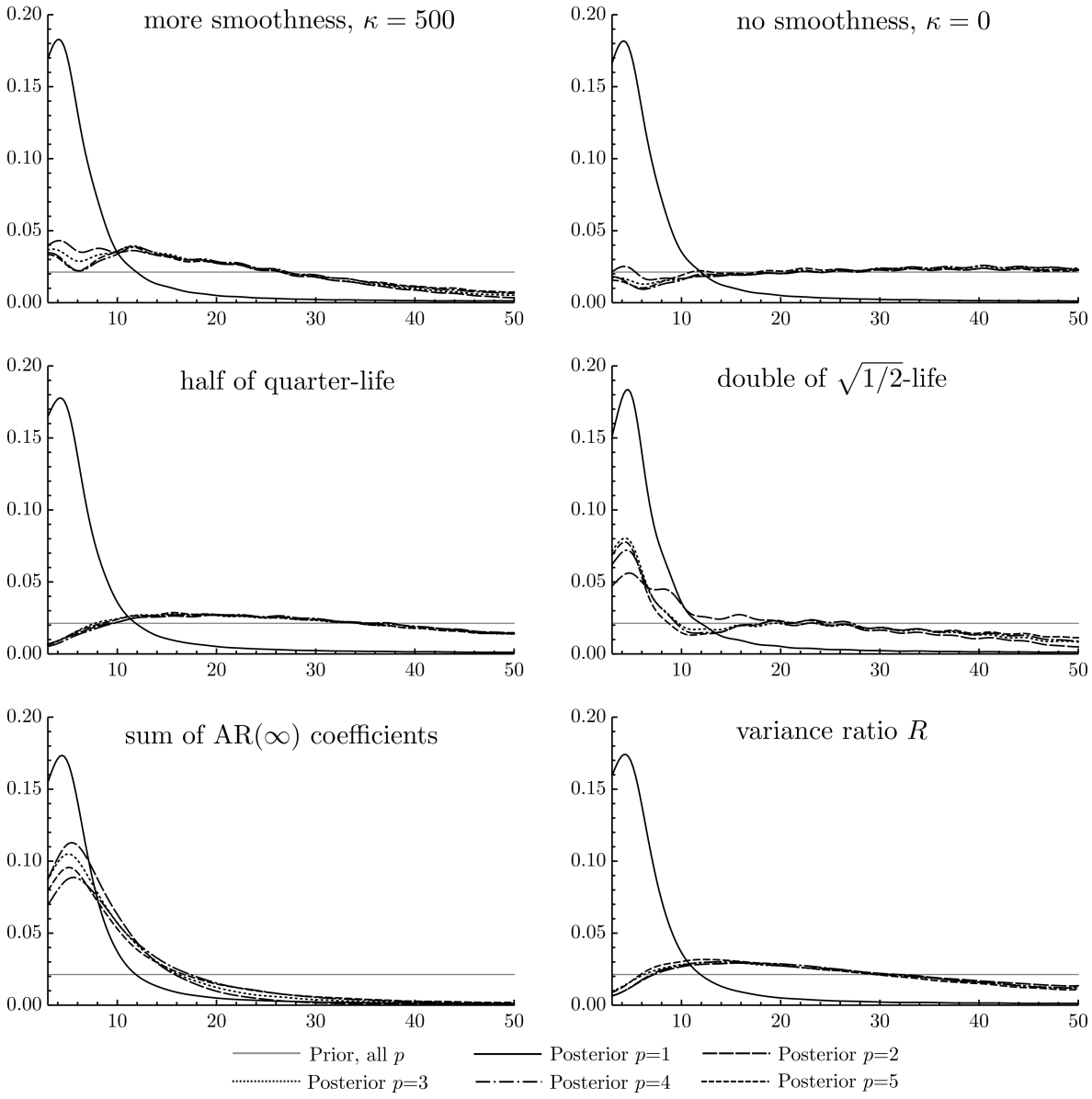
so that for $p = 1$, all four measures coincide.

For each of these six variations, we recompute the prior π_τ so that the resulting prior $\pi(\mathbf{h})$ on the (transformed) alternative persistence measure yields a flat prior on $[3, 50]$ for all p. Figure 4 plots the resulting posterior distribution, in analogy to the right hand side of Figure 3, and Table 3 reports the Bayes factors relative to the $p = 1$ model. At least qualitatively, there is only modest sensitivity to either the smoothness parameter κ or the definition of the half-life in the sense that in all variations, models with $p > 1$ are strongly preferred by the data, and they indicate the presence of substantially more persistence compared to the $p = 1$ specification.

A.6 Frequentist Test of $H_0 : p = 1$

In the context of a GLTU(p) model, consider a frequentist test of $H_0 : p = p_0$ against $H_1 : p = p_1 > p_0$ based on the observation $\mathbf{x}_T$ in the notation of Section A.4.1. Under the

Figure 4: Prior and Posterior Half-Lives as Function of p under Six Variants



approximation of Corollary 1, $T^{-1/2}(\mathbf{x}_T - \mu\boldsymbol{\iota}) \sim \mathcal{N}(0, \omega^2\boldsymbol{\Sigma}_N)$, where $\boldsymbol{\Sigma}_N$ depends on the GLTU parameters. The hypothesis testing problem thus amounts to inference about the covariance matrix of a $N \times 1$ normal vector.

Restrict attention to tests that are invariant to the transformations $\mathbf{x}_T \mapsto s\mathbf{x}_T + m\boldsymbol{\iota}$ for $s > 0$ and $m \in \mathbb{R}$. All invariant tests can be written as a function of the maximal invariant $(\mathbf{x}_T - \boldsymbol{\iota}\boldsymbol{\iota}'\mathbf{x}_T/N)/||\mathbf{x}_T - \boldsymbol{\iota}\boldsymbol{\iota}'\mathbf{x}_T/N||$. By a standard calculation (cf. King (1980)), the density of this maximal invariant is proportional to (54), which does not depend on ω^2 or μ. The invariant testing problem is thus characterized by $2p_0 - 1$ nuisance parameters under the null hypothesis, and by $2p_1 - 1$ nuisance parameters under the alternative.

Given the non-standard nature of this testing problem, we employ the numerical algorithm of Elliott, Müller, and Watson (2015) to obtain a nearly weighted average power maximizing test. We set the weighting function equal to the prior distribution π under p_1 from the main text, which we numerically approximate by computing averages over 500 independent draws from π. We exclusively consider the case $p_0 = 1$, that is, a specification test of the standard LTU model, so that there is a single scalar nuisance parameter $c > 0$ under the null hypothesis (allowing for $p_0 > 1$ results in a considerably harder computational problem, especially for large N). We restrict the null parameter space for c to equal $(0, 200]$, which rules out very strong mean reversion—$c = 200$ implies a half-life of $(\ln 2)/200 = 0.35\%$ of the sample.

We determine tests for $N = 50$ at the 1% and 5% level against $p_1 = 2$ and $p_1 = 3$, and apply them to the real exchange rate data of Section 5.2. We find that both tests reject at the 1% level, corroborating the Bayes factor results that also favor $p > 1$.

References

Andrews, D. W. K., and H. Chen (1994): "Approximately Median-Unbiased Estimation of Autoregressive Models," *Journal of Business and Economic Statistics*, 12, 187–204.

Bergstrom, A. R. (1985): "The estimation of parameters in non-stationary higher-order continuous-time dynamic models," *Econometric Theory*, 1, 369–385.

Bobkoski, M. J. (1983): "Hypothesis Testing in Nonstationary Time Series," *unpublished Ph.D. thesis, Department of Statistics, University of Wisconsin.*

Brand, L. (1964): "The Companion Matrix and Its Properties," *The American Mathematical Monthly*, 71, 629–634.

Brockwell, P. J. (2001): "Continuous-time ARMA processes," in *Handbook of Statistics 19; Stochastic Processes: Theory and Methods*, ed. by D. N. Shanbhag, and C. R. Rao, vol. 19, pp. 249–276. Elsevier.

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

Campbell, J. Y., and M. Yogo (2006): "Efficient Tests of Stock Return Predictability," *Journal of Financial Economics*, 81, 27–60.

Cavanagh, C. L., G. Elliott, and J. H. Stock (1995): "Inference in Models with Nearly Integrated Regressors," *Econometric Theory*, 11, 1131–1147.

Chan, N. H., and C. Z. Wei (1987): "Asymptotic Inference for Nearly Nonstationary AR(1) Processes," *The Annals of Statistics*, 15, 1050–1063.

Elliott, G. (1998): "The Robustness of Cointegration Methods When Regressors Almost Have Unit Roots," *Econometrica*, 66, 149–158.

——— (1999): "Efficient Tests for a Unit Root When the Initial Observation is Drawn From its Unconditional Distribution," *International Economic Review*, 40, 767–783.

Elliott, G., U. K. Müller, and M. W. Watson (2015): "Nearly Optimal Tests When a Nuisance Parameter is Present Under the Null Hypothesis," *Econometrica*, 83, 771–811.

Elliott, G., T. J. Rothenberg, and J. H. Stock (1996): "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, 64, 813–836.

ELLIOTT, G., AND J. H. STOCK (1994): "Inference in Time Series Regression When the Order of Integration of a Regressor is Unknown," *Econometric Theory*, 10, 672–700.

——— (2001): "Confidence Intervals for Autoregressive Coefficients Near One," *Journal of Econometrics*, 103, 155–181.

GOSPODINOV, N. (2004): "Asymptotic Confidence Intervals for Impulse Responses of Near-Integrated Processes," *Econometrics Journal*, 7, 505–527.

IBRAGIMOV, I. A., AND Y. A. ROZANOV (1978): *Gaussian random processes*. Springer Verlag, Berlin.

JANSSON, M., AND M. J. MOREIRA (2006): "Optimal Inference in Regression Models with Nearly Integrated Regressors," *Econometrica*, 74, 681–714.

JONES, R. H. (1981): "Fitting a continuous time autoregression to discrete data," in *Applied time series analysis II*, ed. by D. F. Findley, pp. 651–682. Academic Press, New York.

JONES, R. H., AND K. M. ACKERSON (1990): "Serial correlation in unequally spaced longitudinal data," *Biometrika*, 77, 721–732.

KASPARIS, I., E. ANDREOU, AND P. C. B. PHILLIPS (2015): "Nonparametric predictive regression," *Journal of Econometrics*, 185, 468–494.

KING, M. L. (1980): "Robust Tests for Spherical Symmetry and their Application to Least Squares Regression," *The Annals of Statistics*, 8, 1265–1271.

KOSTAKIS, A., T. MAGDALINOS, AND M. P. STAMATOGIANNIS (2015): "Robust Econometric Inference for Stock Return Predictability," *The Review of Financial Studies*, 28, 1506–1553.

LANNE, M. (2002): "Testing the Predictability of Stock Returns," *The Review of Economics and Statistics*, 84(3), 407–415.

LIEBERMAN, O., AND P. C. B. PHILLIPS (2014): "Norming Rates and Limit Theory for some Time-Varying Coefficient Autoregressions," *Journal of Time Series Analysis*, 35, 592–623.

LIEBERMAN, O., AND P. C. B. PHILLIPS (2017): "A Multivariate Stochastic Unit Root Model with an Application to Derivative Pricing," *Journal of Econometrics*, 196, 99–110.

Lothian, J. R., and M. P. Taylor (1996): "Real Exchange Rate Behavior: The Recent Float from the Perspective of the Past Two Centuries," *Journal of Political Economy*, 104, 488–509.

Lütkepohl, H. (2005): *New Introduction to Multiple Time Series Analysis*. Springer, Berlin.

Magdalinos, T., and P. Phillips (2009): "Econometric inference in the vicinity of unity," *CoFie Working Paper (7), Singapore Management University*.

Meng, X. L., and W. H. Wong (1996): "Simulating Ratios of Normalizing Constants via a Simple Identity: a Theoretical Exploration," *Stastistica Sinica*, 6, 831–860.

Mikusheva, A. (2007): "Uniform Inference in Autoregressive Models," *Econometrica*, 75, 1411–1452.

——— (2012): "One-dimensional inference in autoregressive models with the potential presence of a unit root," *Econometrica*, 80(1), 173–212.

Moon, H. R., and P. C. Phillips (2000): "Estimation of autoregressive roots near unity using panel data," *Econometric Theory*, 16(06), 927–997.

Müller, U. K., and M. W. Watson (2016): "Measuring Uncertainty about Long-Run Predictions," *Review of Economic Studies*, 83.

——— (2017): "Low-Frequency Econometrics," in *Advances in Economics: Eleventh World Congress of the Econometric Society*, ed. by B. Honoré, and L. Samuelson, vol. II, pp. 63–94. Cambridge University Press.

Murray, C. J., and D. H. Papell (2002): "The purchasing power parity persistence paradigm," *Journal of International Economics*, 56, 1–19.

——— (2005): "The purchasing power parity puzzle is worse than you think," *Empirical Economics*, 30(3), 783–790.

Pham-Dinh, T. (1977): "Estimation of parameters of a continuous time Gaussian stationary process with rational spectral density," *Biometrika*, 64, 385–399.

Phillips, A. W. (1959): "The estimation of parameters in systems of stochastic differential equations," *Biometrika*, 46, 67–76.

Phillips, P. C. B. (1987): "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74, 535–547.

——— (1988): "Regression Theory for Near-Integrated Time Series," *Econometrica*, 56, 1021–1043.

——— (1998): "Impulse Response and Forecast Error Variance Asymptotics in Nonstationary VARs," *Journal of Econometrics*, 83, 21–56.

——— (2014): "On Confidence Intervals for Autoregressive Roots and Predictive Regression," *Econometrica*, 82, 1177–1195.

Robinson, P. M. (2003): "Long-Memory Time Series," in *Time Series with Long Memory*, ed. by P. M. Robinson, pp. 4–32. Oxford University Press, Oxford.

Rossi, B. (2005): "Confidence Intervals for Half-Life Deviations from Purchasing Power Parity," *Journal of Business and Economic Statistics*, 23, 432–442.

Stein, E., and R. Shakarchi (2005): *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, Princeton Lectures in Analysis. Princeton University Press, Princeton.

Stock, J. H. (1991): "Confidence Intervals for the Largest Autoregressive Root in U.S. Macroeconomic Time Series," *Journal of Monetary Economics*, 28, 435–459.

——— (1996): "VAR, Error Correction and Pretest Forecasts at Long Horizons," *Oxford Bulletin of Economics and Statistics*, 58, 685–701.

Stock, J. H., and M. W. Watson (1996): "Confidence Sets in Regressions with Highly Serially Correlated Regressors," *Working Paper, Harvard University*.

Taylor, M. P. (2003): "Purchasing power parity," *Review of International Economics*, 11(3), 436–452.

Torous, W., R. Valkanov, and S. Yan (2004): "On predicting stock returns with nearly integrated explanatory variables," *The Journal of Business*, 77(4), 937–966.

Valkanov, R. (2003): "Long-Horizon Regressions: Theoretical Results and Applications," *Journal of Financial Economics*, 68, 201–232.

Wright, J. H. (2000a): "Confidence Intervals for Univariate Impulse Responses with a Near Unit Root," *Journal of Business and Economics Statistics*, 18, 368–373.

——— (2000b): "Confidence Sets for Cointegrating Coefficients Based on Stationarity Tests," *Journal of Business and Economic Statistics*, 18, 211–222.