Beyond co-integration: new tools for inference on co-movements*

Karim M. Abadir[†‡] and Gabriel Talmain[§]

Abstract: Macroeconomic and aggregate financial series were shown empirically to share an unconventional form of cyclical and persistent dynamics, whose functional form was obtained from the solution of general-equilibrium models with heterogeneous firms. The econometric modelling of equations that link such series requires a new methodology, as existing parametric techniques can cause paradoxical regression results and omit predictabilities. We provide a solution to disentangle the genuine relation between variables (the parameters linking them) from the unconventional dynamics that drive them.

As an application, we show that GBP-USD forward premia have no predictive power for excess returns over 1976–2015 (thus solving this forward-premium puzzle) once the unconventional dynamics of spot rates are modelled. Taking advantage of these dynamics, we uncover a trading strategy which consistently outperforms existing ones in the out-of-sample period 2015–2021, delivering almost treble their profits and yielding a Sharpe ratio of 85%. Hence, even in this heavily traded market, the Efficient Market Hypothesis has been failing for over 45 years as persistent profit opportunities remained unexploited

*We thank the Editor (Fabio Trojani), Associate Editor, and referees for their constructive comments. We also thank seminar participants at universities, central banks, and conferences where this paper was invited.

[†]Department of Economics, American University in Cairo, AUC Avenue, P.O.Box 74, New Cairo 11835, Egypt.

[‡]Business School, Imperial College London, London SW7 2AZ, UK, e-mail: k.m.abadir@imperial.ac.uk.

[§]Department of Economics, Adam Smith Building, 40 Bute Gardens, University of Glasgow, Glasgow G12 8RT, UK, e-mail: Gabriel.Talmain@glasgow.ac.uk.

because of the unconventional dynamics of the spot rate.

Key words: Long memory; Persistent cycles; Auto-Correlation Functions; Co-movements; Macroeconomic and aggregate financial series; Micro-founded general equilibrium; Uncovered Interest Parity; Efficient Market Hypothesis; Exchange rates; Forward-premium puzzle; Currency trading strategies

JEL classification: C32, C51, E37, F31, G10

1 Introduction

There are a number of puzzles where some intuitive and theory-consistent economic relation between macroeconomic or financial variables seems to be violated, with the estimated coefficients of the relation defying economic logic. Many of these paradoxes involve variables whose dynamics are notoriously difficult to model. In addition, in macroeconomics, existing models have been accused of failing to predict the turning points of the economic cycle and the troubles that have followed. Our paper aims to introduce new tools for handling relations between such series in a way that prevents misleading estimates and improves predictability.

The persistence of a series can be depicted by its Auto-Correlation Function (ACF), in addition to its usual time-domain and frequency-domain representations. Motivated by the dynamic solution of the micro-founded general-equilibrium model of Abadir and Talmain (2002), the paper by Abadir, Caggiano, and Talmain (2013) provided the counterpart of Dickey-Fuller unit-root tests for univariate time series in the ACF domain, and it was applied to show that almost all macro series and aggregate indexes fall outside the scope of Auto-Regressive Integrated Moving-Average (ARIMA) and Unit Root (UR) models. The current paper takes it to the next step of modelling the co-movement of such series: if such individual series are not integrated, we need to find an alternative to co-integration

analysis for them. The sheer impact of co-integration on empirical macro and finance shows that a new method of co-movements is needed to re-examine the empirical relations in the literature, after the revelations of the 2013 paper. The need for models incorporating nonlinearity and long memory has been felt in a variety of applications that have led to the introduction of statistical models to address these specific applications; e.g., see Gil-Alana (2001), Guidolin, Hyde, McMillan, and Ono (2009), Kruse (2011), Chauvet and Potter (2013), Kostakis, Magdalinos, and Stamatogiannis (2015), and Chevillon and Mavroeidis (2017).

To give a flavour of the constraints that economic theory imposes on empirical models, we now illustrate why the heterogeneity of firms causes the dynamics of GDP to be non-linear. Denote GDP by Y. It is the summation of the value-added of the firms or sectors in the economy, $Y_1, Y_2, \ldots$, as

$$Y := Y_1 + Y_2 + \cdots = \mathrm{e}^{y_1} + \mathrm{e}^{y_2} + \cdots \neq \mathrm{e}^{y_1+y_2+\cdots}, \tag{1}$$

where the usual logarithmic transformation $y_i := \log Y_i$ $(i = 1, 2, \ldots)$ is needed to model percentage changes in Y_i (it also ensures the dynamics of Y_i remain positive, in addition to variance stabilization). Taking logs, the aggregate $y := \log Y$ satisfies

$$y := \log(\mathrm{e}^{y_1} + \mathrm{e}^{y_2} + \ldots) \neq y_1 + y_2 + \ldots, \tag{2}$$

where the left-hand side entails a highly nonlinear aggregation of the processes for $y_1, y_2, \ldots$ and the two sides are always equal if and only if there is only 1 component in Y (the representative-firm assumption that was disposed of in Abadir and Talmain, 2002).[1] The right-hand side of (2) is the linear aggregation that has been used to generate linear ARIMA processes, including fractionally-integrated I(d) cases which have long memory, but nonlinearity is built into even the simplest national-income accounting as in the equations above. Abadir and Talmain (2002) work out explicitly the analytic solution of their microfounded

[1] Abadir and Talmain (2002) use a generalization of (1), a Dixit-Stiglitz CES aggregator of the output of firms with different technical efficiencies and dynamic adjustments, the components Y_i being dependent because of the interaction of firms through the price system and market clearing.

model, with resulting nonlinear and long-memory dynamics. The intuition is that, with GDP evolving in the long run according to this new process, the interrelation of the variables in the general-equilibrium model implies a shared new form of common stochastic trend and cycle for the solution of the other variables in the system.

A striking message that was subsequently obtained from the graphs in Abadir et al. (2013) is the regularity of the actual ACF compared to the jagged time-paths of these variables, and the implied predictability of key features in the series. These ACFs are very different from the ones implied by finite-order ARIMA models, none of which can generate cycles with such persistence properties, as well as local concavity/convexity features that were found in the ACFs. Furthermore, if there are persistent cycles in the variables, differencing them cannot remove this persistence (unlike in the case of differencing I(1) series).[2] This is why a new econometric methodology is needed to deal with estimating relations between variables containing this type of nonlinear long-memory.

If the dynamics of the variables in a parametric model are not adequately represented, more than just finite-sample efficiency loss can arise, namely biased and inconsistent estimates of the relation linking the variables; e.g., see Maddala and Rao (1973) for an early demonstration in a much simpler context. We exploit the common structure of our ACFs to devise a method to disentangle the co-movements of variables (estimating the parameters of the relation linking them) from the effects of persistence of the individual series.

As an application, we show how our method dramatically reverses a much-debated and long-standing counterintuitive finding in tests of the Uncovered Interest Parity (UIP). For this, we take the longest established and one of the most heavily-traded markets, the GBP-USD foreign exchange (FX) rate, and demonstrate that its subtle dynamics (rather than the forward premium) holds predictive power over currency excess returns. Although our

[2] A simple example is the unemployment rate, a variable that is bounded and yet found in many empirical studies to have a unit root when the model is restricted to the ARIMA class. Removing its persistence is not to be achieved by differencing, and it was demonstrated in the 2013 paper that this series belongs to the new class of processes, rather than the ARIMA class; see also Gil-Alana and Trani (2019).

estimation method below will be more elaborate and will involve a model with multiple variables, here we use a very simple descriptive illustration of these unconventional dynamics in Figure 1, where we plot the ACF of the logarithm of the GBP-USD exchange rate and its fit by our functional form (6) for August 1976 to March 2015.[3] The ACF shows a very persistent cycle but one that decays nevertheless, unlike what ARIMA models can represent. Froot and Thaler (1990, p.188), put their finger on the problem, noting that the forward-premium puzzle could be explained "if only part of this appreciation occurs immediately, and the rest takes some time". We will show that this is precisely the sort of problem that current econometric techniques cannot adequately deal with.

We also show that, on the out-of-sample period April 2015 to April 2021, an FX trading strategy based on our methodology outperforms by a large margin the well-known Carry Trade and Momentum FX strategies, yielding cumulative profits of more than 2.9 times the best competing method. Ours also exhibits a Sharpe ratio of 0.85 against a maximum of 0.25 for the two others. Our strategy uses only data contemporaneously available to the traders to make out-of-sample forecasts, all parameters having been estimated from the preceding period of August 1976 to March 2015, and *not* re-estimated subsequently (hence relying on the stability and robustness of our estimates). This demonstrates simultaneously that the UIP anomaly is not caused by the forward premium, and that the Efficient Market Hypothesis has been failing over the last 45 years even in this ideal currency market where persistent profit opportunities remained unexploited because of the unconventional dynamics of the spot rate. Being able to formulate and model the FX dynamics, through our new methodology, will allow better performance than hitherto.

There has been a growing body of evidence of the existence of more accurate predictability in exchange rates, especially with regard to various key features that our process possesses, leading to a strong argument in favour of our new dynamics. Strong autocorrelations

[3] Such an ACF arises again in Talmain (2018) who extends Abadir and Talmain (2002) to a two-country general-equilibrium model where firms are heterogeneous in each country. This time, the ACF applies to the exchange rate between the two countries' currencies.

have been stressed in Backus, Gregory, and Telmer (1993), Bekaert (1996), and Okunev and White (2003). Evidence of long swings and persistence has been shown in Engel and Hamilton (1990), Diebold, Husted, and Rush (1991), Diebold, Gardeazabal, and Yilmaz (1994), Baillie and Bollerslev (1994, 2000), Maynard and Phillips (2001), Carvalho and Nechio (2011), and Bansal and Shaliastovich (2013). Nonlinearities in the process have also been highlighted in Sarno, Valente, and Leon (2006), Della Corte, Sarno, and Tsiakas (2009), Verdelhan (2010), and Kruse (2011).

The plan is as follows. Section 2 outlines our procedure and explains how it deals with these unconventional dynamics in a single-equation framework. We kept derivations and technical remarks out of Section 2 (and shifted them into the Appendix) to make it widely accessible to users, as this is the section that most applications will require. Section 3 extends the approach to a system of equations. Section 4 introduces the UIP application and demonstrates the puzzle. Section 5 applies our method to it, solving the puzzle and providing a new trading strategy. Section 6 concludes, and the Appendix follows. We also use the acronyms AT, UR, CT, Mom as shorthands for Abadir and Talmain, Unit Root, Carry Trade, Momentum; respectively.

2 The ACF-based procedure for a single equation

This section contains two parts. First, we introduce informally the need for our ACF-based estimation, then we present our estimation procedure.

2.1 The intuition behind the setup

Consider the generic decomposition

$$y_t = y_t^\dagger + u_t, \quad t = 1, 2, \dots T, \tag{3}$$

where $y_t^\dagger$ represents the time-varying "fundamental value" (in an economic sense) of y_t, while u_t are the residual dynamics of adjustment towards such a value. By definition, u_t

is centered around zero and is mean-reverting as will be specified more explicitly after (7) below; otherwise y_t will not revert to its fundamental value $y_t^\dagger$. Denoting the $T \times 1$ vector of stacked y_t values by $\boldsymbol{y} := (y_1, \ldots, y_T)'$, and similarly for $y_t^\dagger$ and u_t, we write $\boldsymbol{y} = \boldsymbol{y}^\dagger + \boldsymbol{u}$.

Difficulties arise with existing parametric techniques when the residuals u_t are not generated by a finite-order linear ARIMA process,[4] but instead by a more complicated process as the results of Abadir et al. (2013) suggest. First, it distorts estimation and inference.[5] Second, it entails a loss of predictability if the new dynamics exist and are not modelled.

One may wish to think of the special case of $\boldsymbol{y}^\dagger$ being the linear relation $\boldsymbol{y}^\dagger = \boldsymbol{X}\boldsymbol{\beta}$. Our presentation focuses on the nonlinearity in the dynamics of the residuals $\boldsymbol{u}$, rather than on the specifics of the functional form relating $\boldsymbol{y}^\dagger$ to $\boldsymbol{X}$. Naturally, our procedure will estimate simultaneously the parameters of the relation linking $\boldsymbol{y}$ to $\boldsymbol{X}$ (such as $\boldsymbol{\beta}$) as well as the parameters governing the process $\boldsymbol{u}$.

The ACF $\rho_1, \rho_2, \ldots$ of a process $\{u_t\}_{t=1}^T$ is

$$\rho_\tau := \frac{\operatorname{cov}(u_t, u_{t-\tau})}{\sqrt{\operatorname{var}(u_t)\operatorname{var}(u_{t-\tau})}}, \tag{4}$$

where $\rho_0 \equiv 1$. Long memory is a case where this function of τ decays very slowly as τ increases, typically at a hyperbolic speed and hence much slower than the exponential rate of decay obtained for stationary AR models. Unlike unit root models, shocks to a long-memory process do not have an everlasting impact. More details of standard long-memory can be found in Beran (1992) and Robinson (1994). Their origin from aggregation can be found in Robinson (1978), Granger (1980), Chambers (1998), Chevillon, Hecq, and Laurent (2015). By definition, long memory means that the process for $\{u_t\}$ is stationary.

[4] The relevance of the required finite order can be seen in Abadir and Taylor (1999) and Remark A3 in our Appendix below.

[5] Adjusting the Standard Errors (SEs) for omitted autocorrelation has been extended to the case of long-memory that has a spectral singularity at frequency zero; see Abadir, Distaso, and Giraitis (2009) for a comparison of two methods, including the widely-used Newey-West SEs. Although feasible, there has been no extension to the relatively new case of nonzero frequency, which is the case considered here. Furthermore, parameter estimates (not just their SEs) are also affected by omitted long memory.

The autocorrelation matrix of $\boldsymbol{u}$ can be written as

$$\boldsymbol{R} := \begin{pmatrix} 1 & \rho_1 & \rho_2 & \ddots & \rho_{T-2} & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \ddots & \ddots & \rho_{T-2} \\ \rho_2 & \rho_1 & \ddots & \ddots & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \rho_1 & \rho_2 \\ \rho_{T-2} & \ddots & \ddots & \rho_1 & 1 & \rho_1 \\ \rho_{T-1} & \rho_{T-2} & \ddots & \rho_2 & \rho_1 & 1 \end{pmatrix}. \tag{5}$$

If one were dealing with the simple case of a stationary AR(1) with autoregressive parameter ρ, one would have had $\rho_\tau = \rho^\tau$ and knowledge of ρ alone would have allowed filling the whole $\boldsymbol{R}$ matrix. But in the general case of (5), estimating $\boldsymbol{R}$ requires estimating $\rho_1, \ldots, \rho_{T-1}$, that is $T-1$ parameters, while only T data points are available. Our solution to this parametrization issue is to let ρ_τ take the functional form in Abadir, et al. (2013)

$$\rho_\tau \approx \frac{1 - a\left[1 - \cos\left(\omega\tau\right)\right]}{1 + \left(b\tau\right)^c} \qquad \left(b, c > 0,\ \omega \in (0, 2\pi)\right), \tag{6}$$

with only 4 parameters to estimate rather than $T-1$.[6] It is assumed that a, b, c, ω combine to produce a positive definite $\boldsymbol{R}$ for any T; see Remark A1 in the Appendix for more details. Fourier inversion of this ACF (6) gives a spectral density $f(\lambda)$ that is proportional to $|\lambda - \omega|^{c-1}$ as $\lambda \to \omega$ and is bounded elsewhere; that is, at frequency ω, there is a singularity when $c \in (0, 1)$. For linear long-memory ARIMA(p, d, q) processes having $d \in (0, \frac{1}{2})$,[7] the spectrum has a singularity at the origin that is proportional to $|\lambda|^{-2d}$, giving the correspondence $c = 1 - 2d$ if $\omega = 0$ but not otherwise. Giraitis, Hidalgo, and Robinson (2001) and Hidalgo (2005) give a frequency-domain method of estimating ω and d.

Abadir et al. (2013) show that this 4-parameter functional form (6) represents the dynamics of almost all known macroeconomic variables more accurately than ARIMA models,

[6] The estimation of autocorrelation matrices such as (5) can be improved, especially in the case of big data, by applying a flat-top kernel to the sample autocorrelation; see McMurry and Politis (2010) and Jentsch and Politis (2015).

[7] In fractional I(d), long-memory is the case $d \in (0, \frac{1}{2})$. If a series has more persistence, it is differenced to map the memory parameter from $[\frac{1}{2}, \frac{3}{2})$ to $[-\frac{1}{2}, \frac{1}{2})$.

but their context is that of a single variable in the ACF domain. Here, we deal with incorporating ACFs of this functional form into multivariate time-domain estimation, in order to extract the relation linking the variables.

2.2 Maximum Likelihood procedure

We now present a Maximum Likelihood (ML) procedure to estimate jointly the parameters of $y_t^\dagger$ and those of the ACF of u_t. To simplify the exposition, we adopt the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \text{ with } \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Gamma}), \tag{7}$$

where $\boldsymbol{u}$ has the $T \times T$ autocovariance matrix $\boldsymbol{\Gamma}$, with $\boldsymbol{R}$ its corresponding autocorrelation matrix whose elements are defined by (6), and we assume that $\boldsymbol{\Gamma}$ is proportional to the autocorrelation matrix $\boldsymbol{R}$ in (5)–(6). The proportionality assumption can be relaxed, as explained after (A2) in the Appendix. We now give a method to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ jointly.

We adopt the conditions of Robinson and Hidalgo (1997) which are sufficient (but not necessary) for the asymptotic normality below to hold. They allow for the regressors $\boldsymbol{X}$ to also have long memory. Essentially, we will show in Proposition 1 that our objective function boils down to a concentrated likelihood for estimating the autocorrelation matrix $\boldsymbol{R}$, after which the distributional results are known and limiting normal in Proposition 2.

Denote the determinant of a square matrix $\boldsymbol{M}$ by $|\boldsymbol{M}|$ and, for any given $\boldsymbol{R}$, define

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} := \left(\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \tag{8}$$

as a function of $\boldsymbol{R}$. Then, we have the following result for our ML Estimator (MLE) of $\boldsymbol{R}$ and $\boldsymbol{\beta}$.

Proposition 1 *The MLE of $\boldsymbol{R}$, denoted by $\widehat{\boldsymbol{R}}$, is obtained by maximizing the concentrated log-likelihood (up to constant and scale of $\frac{1}{2}$)*

$$\psi := -\log\left|\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}\right)'\boldsymbol{R}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}\right)\boldsymbol{R}\right| \tag{9}$$

with respect to the four parameters of the ACF given in (6), and the corresponding MLE of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{R}}}$.

The objective function to be maximized is nonlinear in $\boldsymbol{R}$, and a grid search over the 4 parameters of the ACF may be needed to ensure that a global maximum is achieved. The optimization of the joint likelihood (for $\boldsymbol{R}$ and $\boldsymbol{\beta}$) now depends on only 4 parameters that determine the whole autocorrelation matrix $\boldsymbol{R}$. The parsimonious parameterization of the autocorrelation matrix $\boldsymbol{R}$ has allowed us to build our estimation procedure. This parameterization transcends the chosen estimation method, and it can be used as an input for methods other than ML. We chose ML because of its statistical optimality properties, but other choices are now feasible and can employ our parameterization of $\boldsymbol{R}$. One such additional method is given in the Appendix, where we also give in (A4) the estimator of the remaining scale parameter σ^2 needed for the asymptotic variance that makes $\boldsymbol{\Gamma}$ proportional to $\boldsymbol{R}$; e.g., see (A8).

Our MLE is asymptotically valid as a pseudo or quasi MLE under more general conditions than normality in (7), which cover a wide range of distributions of $\boldsymbol{u}$. For this to hold more generally in a likelihood-based estimation procedure, we need to assume regularity conditions with respect to the density from which the sample $\boldsymbol{u}$ would be drawn in (7), and these are stated in the Appendix.

Our MLE satisfies the following asymptotic result.

Proposition 2 *The estimator $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically normal for any $c > 0$. Furthermore, when $\mathbf{X}$ contains no deterministic trends, the limiting distribution of $\sqrt{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is a non-degenerate normal with mean $\mathbf{0}$ and finite variance matrix.*

Note that the consistency rate of $\widehat{\boldsymbol{\beta}}$ will depend on the presence of deterministic components in $\boldsymbol{X}$, but that normalized statistics such as likelihood ratios and t-ratios do not require these rates.[8] For illustrations, see Section 5 below.

[8]Note also that if estimates of $\boldsymbol{\Gamma}$ are substituted into a test statistic for $\boldsymbol{\beta}$, a projection-type adjustment needs to be made to the variance used in constructing the test statistic if it is of the Wald type (such as t-ratios); see Pierce (1982) for details of this adjustment.

3 The ACF-based procedure for a system of equations

We will write the system in its general reduced-form specification. Let us define

$$\boldsymbol{y} := \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_k \end{pmatrix}, \quad \boldsymbol{z} := \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_m \end{pmatrix}, \tag{10}$$

where each of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ contains the T observations on the k endogenous series, and the m series in $\boldsymbol{z}$ are allowed to include lagged dependent variables. The Vector Auto-Regression (VAR) with exogenous variables is a special case of the reduced-form

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_k \end{pmatrix} = (\boldsymbol{B} \otimes \boldsymbol{I}_T) \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_m \end{pmatrix} + \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_k \end{pmatrix}, \tag{11}$$

where the last vector is to be written as $\boldsymbol{u}$, the matrix $\boldsymbol{B}$ is $k \times m$, and $\otimes$ denotes the Kronecker product.

Let $\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_k$ denote the k univariate autocovariance matrices (each of size $T \times T$) of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$, and $\boldsymbol{\Sigma}$ denote the $k \times k$ variance matrix that generalizes σ^2 of (A2). Write the T-dimensional vectors $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_k$ for the k ACF-transformed residuals (one for each endogenous variable) that have no autocorrelation in their marginal distributions $\mathrm{N}(\boldsymbol{0}_T, \sigma_i^2 \boldsymbol{I}_T)$ for $i = 1, \ldots, k$.[9] Then, letting $\boldsymbol{\Gamma}_i = \sigma_i^2 \boldsymbol{L}_i \boldsymbol{L}_i'$, the model can be written as

$$\boldsymbol{L}^{-1} \boldsymbol{y} = \boldsymbol{L}^{-1} (\boldsymbol{B} \otimes \boldsymbol{I}_T) \boldsymbol{z} + \boldsymbol{\varepsilon}, \tag{12}$$

where

$$\boldsymbol{L} := \mathrm{diag}(\boldsymbol{L}_1, \ldots, \boldsymbol{L}_k),$$

with diag denoting a block-diagonal matrix, and $\mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_T$ (hence $\mathrm{var}(\boldsymbol{u}) = \boldsymbol{L} (\boldsymbol{\Sigma} \otimes \boldsymbol{I}_T) \boldsymbol{L}'$). Notice that ε_{it} is an uncorrelated sequence over time but contemporaneously correlated with ε_{jt} $(i, j = 1, \ldots, k)$, and these are transformed such that each

[9] Note that these k ACF-transformed errors have covariances σ_{ij} from the off-diagonal terms of $\boldsymbol{\Sigma}$. Also, we use the convention $\sigma_{ii} \equiv \sigma_i^2$.

resulting u_{it} has its own nonlinear long-memory dynamics of the previous section, with its own set of ACF parameters from $\boldsymbol{\Gamma}_i$, while at the same time being contemporaneously correlated with u_{jt} for general $\boldsymbol{\Sigma}$.

The typical block of $\boldsymbol{L}^{-1}(\boldsymbol{B} \otimes \boldsymbol{I}_T)$ is $\beta_{ij}\boldsymbol{L}_i^{-1}$: the i-th T-dimensional block in the rows of (11) is transformed by $\boldsymbol{L}_i^{-1}$ to remove the autocorrelation of $\boldsymbol{u}_i$, and can be written as

$$\boldsymbol{L}_i^{-1}\boldsymbol{y}_i = \boldsymbol{L}_i^{-1}\sum_{j=1}^{m}\beta_{ij}\boldsymbol{z}_j + \boldsymbol{\varepsilon}_i \equiv \widetilde{\boldsymbol{X}}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \tag{13}$$

where $\widetilde{\boldsymbol{X}}_i := (\boldsymbol{L}_i^{-1}\boldsymbol{z}_1, \ldots, \boldsymbol{L}_i^{-1}\boldsymbol{z}_m)$ and $\boldsymbol{\beta}_i := (\beta_{i1}, \ldots, \beta_{im})'$. This allows us to reformulate the model in a less compact form that will be needed to simplify the estimation procedure.[10] It is the format used by Zellner (1962) for Seemingly Unrelated Regression Equations (SURE),

$$\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{14}$$

where

$$\widetilde{\boldsymbol{y}} := \begin{pmatrix} \boldsymbol{L}_1^{-1}\boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{L}_k^{-1}\boldsymbol{y}_k \end{pmatrix}, \quad \widetilde{\boldsymbol{X}} := \operatorname{diag}(\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_k), \quad \boldsymbol{\beta} := \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_k \end{pmatrix}.$$

By

$$\log|\boldsymbol{L}(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_T)\boldsymbol{L}'| = T\log|\boldsymbol{\Sigma}| + \sum_{i=1}^{k}\log|\boldsymbol{L}_i\boldsymbol{L}_i'| = T\log|\boldsymbol{\Sigma}| - \sum_{i=1}^{k}\left(T\log\sigma_i^2 - \log|\boldsymbol{\Gamma}_i|\right),$$

the normal log-likelihood in terms of the model's $\boldsymbol{u}$ is obtainable from (12) then reformulated via $\boldsymbol{\varepsilon} = \boldsymbol{L}^{-1}\boldsymbol{u}$ and (14) as

$$\begin{aligned}\psi :=& -T\log|\boldsymbol{\Sigma}| + \sum_{i=1}^{k}\left(T\log\sigma_i^2 - \log|\boldsymbol{\Gamma}_i|\right) - \boldsymbol{u}'\boldsymbol{L}^{-1\prime}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_T)\boldsymbol{L}^{-1}\boldsymbol{u} \\ =& -T\log\left(\frac{|\boldsymbol{\Sigma}|}{\prod_{i=1}^{k}\sigma_i^2}\right) - \sum_{i=1}^{k}\log|\boldsymbol{\Gamma}_i| - \left(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_T\right)\left(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}\right)\end{aligned} \tag{15}$$

[10] This equation also shows that we have the flexibility to allow the number of right-hand side variables m to vary with each equation i. It would have been more cumbersome to write this in the earlier equations, but now we can easily alter the summation in (13) to $\sum_{j=1}^{m_i}$. This will not affect subsequent theoretical derivations, and it should be implemented in computations to avoid including irrelevant variables.

up to a constant in T and a scale of $\frac{1}{2}$ as in the proof of Proposition 1. Note that $|\boldsymbol{\Sigma}| / \prod_{i=1}^{k} \sigma_i^2 = 1$ (i.e., the first term of ψ drops out) when $k = 1$ (as in the previous section) or when $\boldsymbol{\Sigma}$ is a diagonal matrix (no contemporaneous correlation of $\boldsymbol{\varepsilon}$). Otherwise, this ratio measures the deviation of $\boldsymbol{\Sigma}$ from diagonality and is always ≤ 1 by the inequality of geometric and arithmetic means; see Abadir, Heijmans, and Magnus (2018, p.196).

Our ML estimator for this model is obtained by maximizing ψ. For any given $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma} := (\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_k)$, this function is optimized for unrestricted estimates by

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma},\boldsymbol{\Gamma}} := \left(\widetilde{\boldsymbol{X}}' \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_T \right) \widetilde{\boldsymbol{X}} \right)^{-1} \widetilde{\boldsymbol{X}}' \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_T \right) \widetilde{\boldsymbol{y}}.$$

Our estimation procedure is as follows:

1. Estimate the parameters of $\boldsymbol{\Gamma}$ from the k individual equations, estimating the 4-parameter ACF for each positive-definite $\boldsymbol{\Gamma}_i$ $(i = 1, \ldots k)$ as in the previous section, leading to an estimator that we denote by $\widetilde{\boldsymbol{\Gamma}}$. This enables us to calculate $\widetilde{\boldsymbol{y}}$ and $\widetilde{\boldsymbol{X}}$.

2. We then follow the original SURE approach of estimating the k-dimensional $\boldsymbol{\Sigma}$ by

$$\widehat{\boldsymbol{\Sigma}} := \frac{1}{T} \begin{pmatrix} \widehat{\boldsymbol{\varepsilon}}_1' \\ \vdots \\ \widehat{\boldsymbol{\varepsilon}}_k' \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\varepsilon}}_1 & \ldots & \widehat{\boldsymbol{\varepsilon}}_k \end{pmatrix}$$

   with $\widehat{\boldsymbol{\varepsilon}} := \left(\boldsymbol{I}_{Tk} - \widetilde{\boldsymbol{X}} (\widetilde{\boldsymbol{X}}' \widetilde{\boldsymbol{X}})^{-1} \widetilde{\boldsymbol{X}}' \right) \widetilde{\boldsymbol{y}}$ the residuals from an Ordinary Least Squares (OLS) regression of $\widetilde{\boldsymbol{y}}$ on $\widetilde{\boldsymbol{X}}$. As usual, the information matrix is block-diagonal (orthogonal parametrization) with respect to $\boldsymbol{\Sigma}$, so the remainder of the procedure determines estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ only.

3. We could take the simple estimator $\widetilde{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}},\widetilde{\boldsymbol{\Gamma}}}$. But if $\boldsymbol{\Sigma}$ is not diagonal (which can be tested, with standard inference applying to $\widehat{\boldsymbol{\Sigma}}$), efficient estimation of a system requires us to maximize

$$-\sum_{i=1}^{k} \log |\boldsymbol{\Gamma}_i| - \left(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}} \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}},\boldsymbol{\Gamma}} \right)' \left(\widehat{\boldsymbol{\Sigma}}^{-1} \otimes \boldsymbol{I}_T \right) \left(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}} \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}},\boldsymbol{\Gamma}} \right) \tag{16}$$

with respect to the $4k$ parameters in $\boldsymbol{\Gamma}$ to get $\widehat{\boldsymbol{\Gamma}}$ and hence $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\Sigma}},\widehat{\boldsymbol{\Gamma}}}$.

We can now state the following result to complete the procedure.

Proposition 3 *The estimators $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ are consistent and asymptotically normal for any $c > 0$. Furthermore, when $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ contain no deterministic trends, the limiting distributions of $\sqrt{T}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\sqrt{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ are non-degenerate normals with means $\mathbf{0}$ and finite variance matrices.*

4 Application: Statement of UIP Theorem and Forward-Premium Puzzle

One test for the efficiency of the foreign exchange market, going back to Fisher (1930), is that speculators will equalize the expected return on the similar short term assets across countries once converted to the same currency. However, a large number of authors analyzing the data have found systematic deviations from this norm. The data seemed to lend support to a very substantial negative relation between the future returns on holding a currency and the current forward premium on it. This is known as the forward-premium puzzle or anomaly. Many authors have studied this very counterintuitive result and excellent summaries are found in Froot and Thaler (1990), Lewis (1995), Engel (1996).

This section contains two parts: the three main alternative formulations of a test of the UIP theorem, followed by the empirical puzzle. The first two formulations are the most popular, in terms of excess returns and in terms of currency depreciation, respectively. The latter also provides a bridge to the third form which is in terms of the levels of the variables and shows how the UIP regression can be expressed in terms of our estimation method. Note that Apte, Sercu, and Uppal (2004) recommend using levels in the related context of purchasing power parity. They show that it leads to the model in levels outperforming the other traditional formulations in terms of first-differences and differences between variables.

4.1 Forms of UIP regressions, and potential sources of failure

UIP regressions. Ignoring transaction costs, a US investor's excess return on investing in a one-period GBP-denominated riskless bond over the corresponding USD bond is

$$r_{t+1} := i_t^{£} - i_t + \Delta s_{t+1}, \tag{17}$$

where $i_t^{£}$ and i_t are the logarithmic returns on the UK and US bond, respectively, s_t is the log of the spot exchange rate S_t (the GBP-USD rate, such that £1 is worth $\$S_t$) and $\Delta s_{t+1} := s_{t+1} - s_t$ is the logarithmic rate of depreciation of the US currency. The UIP hypothesis requires that, under symmetry and frictionless assumptions, traders should expect high interest rate currencies to depreciate relative to low interest ones, thus ensuring no systematic excess returns. The Efficient Market Hypothesis adds that r_{t+1} should not be predictable with any information available at time t. In particular, the forward premium $(f_t - s_t)$, where f_t is the log of the forward rate F_t, should have no explanatory power. We briefly consider three essentially-equivalent formulations of a test for this joint hypothesis.

The three main formulations of the UIP regression are:

$$1^{\text{st}} \text{ Form:} \qquad r_{t+1} = \alpha + \beta\,(f_t - s_t) + u_{t+1}, \tag{18}$$

$$2^{\text{nd}} \text{ Form:} \qquad \Delta s_{t+1} = \alpha + \gamma\,(f_t - s_t) + u_{t+1}, \tag{19}$$

$$3^{\text{rd}} \text{ Form:} \qquad s_{t+1} = \alpha + \gamma\,(f_t - s_t) + \theta s_t + u_{t+1}. \tag{20}$$

The first form is the usual direct implementation, where unpredictability is checked by testing $\beta = 0$. The literature has found significantly negative estimates $\widehat{\beta}$ of β, leading to the forward premium puzzle which we shall return to in the next subsection.

The second form of the UIP regression is equivalent to the first one, up to the Covered Interest Parity (CIP) relation $i_t^{£} - i_t = -(f_t - s_t)$.[11] By substituting this CIP relation into

[11] Consider converting \$1 into £$1/S_t$, investing this amount in the foreign bond, and selling forward the forthcoming £$\left(1 + I_t^{£}\right)/S_t$ at the forward rate F_t. Since all of these transactions can be completed today at no risk, the USD yield on this $(i_t^{£} + f_t - s_t)$ must be equal to the USD yield of investing in a domestic bond (i_t), by arbitrage.

(18), and using definition (17), we get the second form of the UIP regression with $\gamma = \beta + 1$. In normal market conditions, and unlike the UIP relation, the CIP holds almost exactly (hence $r_{t+1} \approx s_{t+1} - f_t$ from (17)). It has been argued that the CIP for the Euro-USD pair failed to a significant degree during the 2008 financial crisis. We checked the CIP for the GBP-USD pair. A total of eight violations were found in our sample of 543 observations, with only half of them being of a substantial magnitude.

The conditions for a random walk for $\{s_t\}$ can be seen more clearly in the second form. If $\beta = -1$ hence $\gamma = 0$ (which would violate UIP), the exchange rate follows a random walk if one were to believe that u_{t+1} did not contain further dynamics. (Testing of β and the dynamics of u_{t+1} will follow in Section 5.) Alternatively, a random walk consistent with UIP can arise if $\{f_t - s_t\}$ is an i.i.d. series or, as a degenerate special case of it, $\{f_t - s_t\}$ is approximately constant as assumed in the pricing of currency options.[12]

The third form of the UIP regression is obtained by recalling that $\Delta s_{t+1} := s_{t+1} - s_t$ and adding s_t to both sides of (19), giving $\theta = 1$. The UIP hypothesis is then a test of $\gamma = 1$ and $\theta = 1$. This formulation is in terms of the levels of the variables, with s_{t+1} as the dependent variable and only u_{t+1} is contemporaneous to it in the equation.

Two potential failures of the UIP regressions. The third form highlights the equivalence between forecasting the excess return (r_{t+1}) and the spot rate (s_{t+1}), up to the CIP. It outlines two possible breakdowns of the UIP regression. The coefficient γ of the forward premium may be different from 1 ($\gamma \neq 1$) and/or the spot rate may not be a unit-root process ($\theta \neq 1$). Figure 1 is already hinting that the latter is certainly a source of concern, the ACF being very different from that of a unit-root process; see also Footnote 12.

One of the possibilities we will consider is the following. Investors could exploit the information in the forward premium (or interest-rate differential) and eventually it would lead to no further profit opportunities, but they may be unable to capitalize on the de-

[12] This random walk hypothesis for s_t is negated visually by Figure 1 where we see persistent cycles whose memory dies out eventually. Also, a random walk would imply an increasing variance over time, $\text{var}(s_t) \propto t$, but we know that the GBP-USD exchange rate has been confined so far to $S_t \in [1, 2.5]$ almost always since its floating in 1973.

viations of the spot rate from the random walk because these are much harder to model. In this case, the forward premium will lack explanatory power and lead to $\gamma = 0$, but the coefficient θ of the spot rate will be unconstrained and the remaining unconventional dynamics of s_{t+1} will be left over in u_{t+1} (which we will model explicitly through ACFs). We will refer to this case as the *irrelevance of the forward premium*. Interestingly, Hassan and Mano (2019) established that the estimate of $\widehat{\beta}$ had no statistical significance for the returns on carry trade strategies. Indeed, they find that across-time variations account for a large percentage of the dollar trade anomaly.

4.2 The Forward-Premium Puzzle

We start by presenting the results using traditional methods, to verify the presence of the puzzle in GBP-USD data. We use GBP-USD monthly data from Datastream for the period August 1976 to April 2021. Running the regression (18) on the original data, OLS gives

$$\begin{array}{rcll} \widehat{r}_{t+1} & = & -0.002 & -2.38\ (f_t - s_t)\,, \\ & & (-1.50) & (-3.86) \end{array} \tag{21}$$

where the t-ratios are given in parentheses below the estimates. The hypotheses they test are that the coefficients are 0. They show that the coefficient of $(f_t - s_t)$ is significantly different from its anticipated value of 0. It seems that forward rates violate the UIP in a puzzling way, if one were to believe these estimated parameters.

By the CIP's $f_t - s_t = -(i_t^{\pounds} - i_t)$, (18) becomes

$$r_{t+1} = \alpha - \beta(i_t^{\pounds} - i_t) + u_{t+1} \tag{22}$$

and the estimate $\widehat{\beta} \ll 0$ from (21) means that positive interest differentials are associated, on average, with substantial positive excess returns on the high-yield currency: instead of depreciating as expected from UIP, high-interest currencies puzzlingly outperform one period later. Also, this correlation has been associated in practice with carry trades. But, one should keep in mind the following:

- Correlation measures only the linear part of a relation; e.g., the population (not sample) correlation of a symmetric x with $y := (x - \mathrm{E}(x))^2$ is exactly zero, even though y is (by definition) fully determined by x.

- Correlation measures only association, not true dependence whose genuine source may be other factors as we shall show. We will see that estimating (21) or (22) does not give the best model to fit the data (because of the dynamics in u_{t+1} that have been neglected) and that neither of $(f_t - s_t)$ or $(i_t^{\pounds} - i_t)$ is the optimal predictor of r_{t+1}; the regression is misspecified and the true β is very different from the flawed estimate $\widehat{\beta}$ obtained in this section.

5 Application: Solving the Forward-Premium Puzzle, and a new trading strategy

This section contains two parts. We start with estimation and inference in the UIP regression by means our ACF-based procedure, revealing that the source of the UIP breakdown is not the forward premium and its puzzling coefficient estimate. The estimation is performed on the 464 observations of August 1976 to March 2015,[13] keeping a few initial values back to January 1976 in reserve in case lagged variables need to be considered in the regressions. The remaining 73 observations, April 2015 to April 2021, will be used for out-of-sample forecasting. Our parameter estimates do not rely on any information from the second subsample and, based on our out-of-sample forecast, we devise a trading strategy to show that our methodology outperforms —by a long stretch— the leading existing strategies. Such forecast comparisons could have been performed live after the estimation period.

[13] It is customary to require long samples when estimating long-memory processes, which explains the length of our first subsample. It ends when transitory uncertainty, due to the announcement of the 2016 Brexit referendum, starts to visibly impact the exchange rates and their volatility.

5.1 The new parameter estimates and hypothesis testing

Model with new ACF dynamics in u_t. Using our procedure on the August 1976 to March 2015 subsample, we estimate the dynamics of the spot rate with the model

$$s_{t+1} = \alpha + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \gamma\,(f_t - s_t) + \theta s_t + u_{t+1} \tag{23}$$

with u_t having the ACF dynamics seen in Section 2. The dummies D_1 and D_2 are for the exceptional events of March 1985 (run-up to the Plaza Accord) and October 2008 (Global Financial Crisis), respectively; so $D_{1t} = 1$ when $t =$ March 1985 and $D_{1t} = 0$ otherwise, $D_{2t} = 1$ when $t =$ October 2008 and $D_{2t} = 0$ otherwise. We will refer to this model as the *AT model*. These dummies will also be introduced in the comparison with competing models below, as well as in the baseline UIP model.

First, we estimate the unrestricted version of the regression (23) then, by estimating the restricted versions, we perform Likelihood Ratio (LR) tests of the 3 hypotheses: $\mathrm{H}_0 : \gamma = 1$, $\mathrm{H}_0 : \theta = 1$, and the joint $\mathrm{H}_0 : \gamma = 1$ and $\theta = 1$. We also test the irrelevance of the forward premium as $\mathrm{H}_0 : \gamma = 0$. For the unrestricted regression, we get the joint estimates

$$\widehat{\rho}_\tau = \frac{1 - 1.011\,[1 - \cos\,(0.0565\tau)]}{1 + \tau^{0.124}} \tag{24}$$

and

$$\widehat{s}_{t+1} = 0.14 D_{1t} - 0.14 D_{2t} - 1.25\,(f_t - s_t) + 0.88\,s_t + \widehat{u}_{t+1}, \tag{25}$$

where the estimated constant was insignificant and not reported.[14] There is a slight of notation here: the term $\widehat{u}_{t+1}$ refers to the values of u_{t+1} fitted (or explained) by the ACF whose parameter were estimated in (24). Unlike in standard models, $\widehat{u}_{t+1}$ here is not the observed residual that would make up $\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}}$ whose left-hand side is $\boldsymbol{y}$ (the actual data). We chose this unusual notation in (25) to stress that u_{t+1} contains predictable dynamics that are fitted as part of the model's estimation of s_{t+1}.

We analyzed the unexplained residuals $\widehat{\boldsymbol{\varepsilon}}$ implied by our estimation, to check for model misspecification, as per the discussion after (A2) in the Appendix. They easily passed the

[14] The likelihood surface of our model is flat in combinations of large b and c. To mitigate this problem, we use the normalization $b \leq 1$ for the parameter b that scales the horizontal axis of the ACF.

following diagnostics: AR tests for omitted autocorrelations, Ramsey's RESET test for omitted nonlinearity, and White's two tests for unconditional heteroskedasticity (with and without cross-products of regressors). However, there was some conditional heteroskedasticity left in the residuals, which is usual for this kind of data containing volatility clustering. It would reduce the finite-sample efficiency of the estimates (not a major concern in such a large sample) but not affect their unbiasedness and consistency.

The optimization problem being formulated in terms of the concentrated likelihood, it is straightforward to use LR tests for all four hypotheses which we examine; see also Footnote 8 (following Proposition 2) if one were to use instead t-tests or more generally Wald-type tests. The results are in Table 1, where UR stands for Unit Root, ℓ stands for the log-likelihood ($\ell = \text{const} + \frac{1}{2}\psi$ from Proposition 1 and its derivation in the Appendix),

$$LR := 2\left(\ell_{\text{unrestricted}} - \ell_{\text{restricted}}\right) = \psi_{\text{unrestricted}} - \psi_{\text{restricted}} \overset{a}{\sim} \chi^2(\nu)$$

with degrees of freedom ν equal to the number of restrictions under H_0, and the p-value gives the tail area to the right of any obtained value of LR as $\Pr(\chi^2(\nu) > LR)$.

Model:	unrestricted θ, γ	$\theta = 1$ s has UR $f - s$ free	$\theta = 1,\ \gamma = 1$ UIP	$\gamma = 1$ s free $f - s$ constrained	$\gamma = 0$ s free $f - s$ irrelevant
$\widehat{a}$	1.011	1.005	1.005	1.021	1.014
$\widehat{b}$	1	1	1	1	1
$\widehat{c}$	0.124	0.050	0.050	0.196	0.149
$\widehat{\omega}$	0.0565	0.0658	0.0643	0.0547	0.0555
ℓ	2,102.3	2,093.9	2,090.3	2,098.7	2,101.2
LR		16.71	23.92	7.15	2.14
p-value		0.0%	0.0%	0.7%	14.4%

Table 1: LR tests for alternative models of s_{t+1} in (23) estimated by the AT methodology.

The UIP joint hypothesis $\gamma = 1$ and $\theta = 1$ is heavily rejected again, but the source of the rejection and its implication for the forward premium puzzle are important. Restricting

only $\theta = 1$ (to test this by LR) still leads to rejecting $\theta = 1$, but freeing up the parameter θ of s (to test γ) changes the results substantially: the conclusion of the irrelevance of the forward premium (hypothesis $\gamma = 0$) cannot be rejected even at the 14% level. This coincides with the findings of Hassan and Mano (2019) about the lack of a link between $(f_t - s_t)$ and the performance of carry trades. The results are effectively telling us that s_{t+1} is predictable by its own dynamics, rather than by the forward premium which has been blamed for the UIP failure. Going further, we pin down the culprit for this puzzle, and it is the exchange rates which follow unconventional persistent cycles that are different from a random walk.

In addition, recursive parameter estimates (as the sample is gradually increased until March 2015) indicate the stability of our estimates over time: there are no recurring structural breaks in the model. The instability of the coefficient of $(f_t - s_t)$ in the standard UIP regression, noted by some authors such as Baillie and Bollerslev (2000), has vanished. In Figure 2a, each central line presents recursive parameter estimates for the coefficients of the forward premium $(f_t - s_t)$ and of the spot rate s_t in equation (25), as the sample is increased to its full estimation size. The bands around the estimates are the traditional formula for an approximate ± 2 Standard Error (SE) and corresponding 95% confidence interval. For stability of the parameter estimates, the central lines should become nearly horizontal as the sample is recursively increased to its full size for estimation. Apart from the initial estimates which are based on a handful of data points, both central lines are near horizontal for our estimates and are well within the terminal (full estimation sample) ± 2 SE bands.

Notice that the power of τ in (24) indicates that the memory in u_t decays eventually (unlike in unit-root models), but it does so more slowly than stationary linear AR models can allow. Also notice that the studies finding long memory or strong autocorrelations, which we mentioned in the introduction, are corroborated here and the role of cyclical long-memory in causing the puzzle is explained. In addition, the cyclicality that we estimate explains why linear ARIMA processes need to be augmented with a persistent time-varying

component (risk premium) to explain the data, as in Baillie and Bollerslev (2000), Engel and West (2005), and Engel (2016). Of course, we could improve our results further by including risk premia, transaction costs, and/or peso problems; but this is not the purpose of our analysis.

The NoD model. For a fair comparison, let us look at the UIP regression (25) under the erroneous assumption that the residuals are white noise but adding our two dummies. We will refer to this model as the *No Dynamics* (NoD) model because it does not model the dynamics in u_{t+1}. The results are:

$$\begin{array}{lcccccc} \widehat{s}_{t+1} & = & 0.15D_{1t} & -0.15D_{2t} & -1.18\,(f_t - s_t) & +0.98\,s_t, & \\ \text{(from 0)} & & (5.5) & (-5.6) & (-2.0) & (97.5) & (26) \\ \text{(from 1)} & & & & (-3.8) & (-1.7) & \end{array}$$

where the estimated constant was insignificant and not reported. The coefficient of s_t jumps from 0.88 to almost 1 (a unit root) to pick up the leftover dynamics of u_{t+1} which have been omitted from the model. The t-ratios for significance from 0 or 1 are in the two lines below the parameters estimates, respectively. Bearing in mind that the model is misspecified and so is inference based on it, the coefficient of the forward premium seems to indicate a marginal significance (from 0) and much more so from the 1 needed for UIP.

Recursive estimation of the parameters in Figure 2b contrast with the earlier results of Figure 2a, as we now see coefficients trending up then down or vice versa, something that was not present in the flatter recursive estimates for AT. The new results indicate that the coefficient of $(f_t - s_t)$ was significantly different from zero for most of the sample period, notably for January 1985 to February 2009 and towards the end of 2014, unlike for our estimates. The coefficient for the autoregressive component increases into explosive territory, before reverting to being a unit root. As noted by a referee, this makes the SE bands for θ indicative rather than asymptotically accurate (if the model were correctly specified and nonstationarity existed), because the distribution theory would be different under a unit or explosive root.

For completeness, Table 2 presents the results of the corresponding χ^2 tests of this

incomplete (hence misspecified) model. They lead to the clear rejection of UIP again, but this time erroneously implying a unit root because the full dynamics of s_{t+1} (through u_{t+1}) were not considered.

Model:	unrestricted	$\theta = 1$	$\theta = 1,\ \gamma = 1$	$\gamma = 1$	$\gamma = 0$
ℓ	$1,016.1$	$1,014.7$	$1,008.0$	$1,009.1$	$1,014.0$
LR		2.76	16.22	14.08	4.16
standard p-value		9.7%	0.0%	0.0%	4.1%

Table 2: LR tests for alternative models of s_{t+1} in (23) under the NoD.

But could the statistical gains from uncovering our dynamics be deceiving? Is there some economic gain that can be built on the statistical one? If our results on the predictability of the spot rate are useful, then they should deliver an FX strategy that would yield economically-significant profits. We now investigate this.

5.2 Forecasting and FX trading strategy

Our methodology has identified the source of failure in the UIP regressions, and has provided a constructive approach to modelling s_{t+1} hence r_{t+1}. There is therefore some predictability in s_{t+1}, and some profit-making investment strategy should exist. Such a strategy should be able to outperform the well-known FX strategies used by traders such as Carry Trade (CT) or Momentum (Mom). Can our methodology help us identify one?

As our goal is to compare the profitability of different strategies, not to see how this profitability can be refined and optimized for actual trading, we keep the trading strategies at their simplest. We ignore all trading costs and other market frictions. We consider a single currency pair, the well-established Cable (GBP-USD), instead of a panel of currencies. We cumulate but do not compound the gains or weigh the amount invested by, say, the extent of predicted excess return which would favour our method. We only look at 1-step-ahead trading although, like co-integration, our model's strength is to reveal the long-run relations. Even so, we will see that we can achieve substantial gains.

For each trading strategy $\mathcal{S} \in \{\text{AT, CT, Mom, NoD}\}$, we estimate the 1-step-ahead forecast for the spot rate $\widehat{s}_{t+1}^{\mathcal{S}}$ for April 2015 to April 2021. The forecasted first difference and excess return are easily calculated as

$$\Delta\widehat{s}_{t+1}^{\mathcal{S}} := \widehat{s}_{t+1}^{\mathcal{S}} - s_t \text{ and } \widehat{r}_{t+1}^{\mathcal{S}} := \Delta\widehat{s}_{t+1}^{\mathcal{S}} + i_t^{\pounds} - i_t = \Delta\widehat{s}_{t+1}^{\mathcal{S}} - (f_t - s_t)\,.$$

All $\mathcal{S}$'s are margin strategies:

1. At time t, if $\widehat{r}_{t+1}^{\mathcal{S}} > 0$, the strategy $\mathcal{S}$ says "buy". The trader borrows an amount \$1 in the domestic currency and uses it (after converting into the foreign currency) to buy the foreign asset. If $\widehat{r}_{t+1}^{\mathcal{S}} < 0$, $\mathcal{S}$ says "sell". The trader borrows \$1 worth in the foreign market and (after converting it) buys \$1 of the domestic asset. The case $\widehat{r}_{t+1}^{\mathcal{S}} = 0$ has a zero probability, as r_t has a continuous distribution, and it is not observed in our study.

2. In the next period, the actual excess return, r_{t+1}, is revealed. The trader collects the payoff in the currency chosen, and uses the proceeds to pay back the loan in the other currency. If the trader has correctly predicted the sign of the excess return (i.e., $\text{sgn}\left(\widehat{r}_{t+1}^{\mathcal{S}}\right) = \text{sgn}\left(r_{t+1}\right)$) the outcome is called a "hit" and the trader's profit is $|r_{t+1}|$. Otherwise, the outcome is called a "miss", and the trader's profit is $-|r_{t+1}|$. Denote this binary outcome as $x_{t+1}^{\mathcal{S}}$ and the profit as $\pi_{t+1}^{\mathcal{S}}$:

$$x_{t+1}^{\mathcal{S}} := \begin{cases} 1 \text{ if "hit" (correct prediction of the sign of } r_{t+1}) \\ 0 \text{ if "miss" (incorrect prediction of the sign of } r_{t+1}) \end{cases}$$

$$\text{and } \pi_{t+1}^{\mathcal{S}} := \begin{cases} |r_{t+1}| \text{ if } x_{t+1}^{\mathcal{S}} = 1 \\ -|r_{t+1}| \text{ if } x_{t+1}^{\mathcal{S}} = 0. \end{cases}$$

3. The cumulated profits of the strategy are $\Pi^{\mathcal{S}} := \sum_{t=0}^{T-1} \pi_{t+1}^{\mathcal{S}}$, and the number of hits $X^{\mathcal{S}} := \sum_{t=0}^{T-1} x_{t+1}^{\mathcal{S}}$. We will also calculate the Sharpe ratios of the strategies.

The 1-step-ahead forecast procedure detailed at the end of the Appendix describes how $\widehat{s}_{t+1}^{\text{AT}}$ if forecasted by AT. The forecasting exercise is based on the post-estimation

subsample where we condition on the ACF parameters estimated on the earlier subsample. It is implicitly a check for whether the ACF still does well after the estimation sample has ended.

The parameters of the ACF and the coefficients of the regression are those already estimated under the null of $\mathrm{H}_0 : \gamma = 0$ of Table 1 and

$$\widehat{s}_{t+1} = 0.06 + 0.15D_{1t} - 0.14D_{2t} + 0.88\, s_t + \widehat{u}_{t+1} \tag{27}$$

instead of (25) where $\gamma = 0$ was not imposed. Note that our stable recursive estimates (see Figure 2a) leading to $\widehat{\theta} = 0.88$ imply that we expect this number to be stable out-of-sample too, unlike recursive estimates that are trending or cycling as more observations are added. For CT, $\widehat{r}^{\mathrm{CT}}_{t+1} \propto i^{f}_{t} - i_t = -(f_t - s_t)$. For Mom, $\Delta\widehat{s}^{\mathrm{Mom}}_{t+1} \propto \Delta s_t$ and, given how small the interest differential was during the forecast period, the fluctuations in the currency dominate and we take $\widehat{r}^{\mathrm{Mom}}_{t+1} \propto \Delta s_t$. Finally, for NoD, $\widehat{s}^{\mathrm{NoD}}_{t+1}$ is estimated from (26).

First we visualize the performance of our out-of-sample forecast for $\widehat{s}^{\mathrm{AT}}_{t+1}$ in Figure 3. Our estimate $\widehat{\omega} = 0.0555$ corresponds to a 9.5 year cycle and can be expected to detect features around this frequency in the spot rate. Indeed, we observe that our forecast picked up the downward trend in the 2015–16 period, and the subsequent upward trend but with a delay of a few months. Clearly, more work is needed to account for the shorter-term dynamics and, potentially, improve the profitability of this strategy. As it stands, our model is mainly concerned with uncovering the long-memory cycle and its implications on estimation, using just one lag of s_t in (27) when more than one lag is needed if one wishes to add a shorter-term cycle (to arise from complex-conjugate AR characteristic roots). However, full modelling of the exchange rate is not the purpose here, rather a comparison of the basic implementation with existing methods.

Table 3 quantifies the success of these strategies in terms of hits and misses over the 73 out-of-sample forecasts of AT. As a benchmark, the first column reports the performance of a trader with Perfect Foresight (PF), i.e., one who could predict s_{t+1} with 100% accuracy, thus providing an upper limit for this class of trading. AT is right more frequently than Carry Trade or Momentum, but is the difference significant? One of the most robust meth-

ods available to test whether two paired series were generated by a single data-generating process is the sign test. We reject at 3.5% probability level the null hypothesis that the median of X for CT is equal to the median for AT, and at the 6.9% level that the median of X for Mom is equal to the one for AT.

	PF	AT	CT	Mom	NoD
number of hits $X^{\mathcal{S}}$	73	43	35	35	35
probability of a hit	100%	59%	48%	48%	48%
p-value of sign test versus AT	-	-	3.5%	6.9%	-
p-value of sign test versus NoD	-	-	50%	50%	-
cumulated profits $\Pi^{\mathcal{S}}$	1.61	0.47	0.16	-0.10	-0.13
Sharpe ratio (annual)	6.92	0.85	0.25	-0.20	-0.19

Table 3: Market performances of AT and other trading strategies.

Hence AT is significantly better at predicting the sign of the excess return, but perhaps it is only better for the smaller excess returns and is beaten badly for the more lucrative large ones? The AT strategy delivers cumulated profits of 0.47, which are 289% better than the next best strategy, CT at 0.16, and achieves 29% of the theoretical maximum PF at 1.61. AT is also a much more consistent strategy with a Sharpe ratio of 0.85, more than treble the next best (CT). Figure 4 plots the cumulative profits for the three strategies over time. The AT strategy uniformly dominates the other strategies throughout the period, not just in its final cumulated profits. Finally, the strategy based on the NoD provides a performance comparable to Momentum, and is much worse than AT.

In Table 4, we analyze further the performance by breaking it down into annual profits and standard deviations. We start the year with April, so that we have 6 full years of comparison instead of only 5 full calendar years. The results show that AT profits uniformly dominate CT and dominate Mom 4 times to 2 (one of the latter is only marginally). This, plus the overall dominance across the whole period, are strong arguments that a generic trader will prefer AT regardless of her/his risk aversion. The standard deviations are pretty similar across methods, year-by-year, because the payoffs in each month are equal in

absolute value. However, there is only one year (2019-20) in which AT's standard deviation is not the lowest (Mom is lower in that year).

	Profits			Standard deviations		
	AT	CT	Mom	AT	CT	Mom
Apr 15 - Mar 16	0.15	0.08	-0.06	0.025	0.027	0.027
Apr 16 - Mar 17	0.15	0.15	-0.11	0.036	0.036	0.037
Apr 17 - Mar 18	-0.12	-0.12	-0.10	0.024	0.024	0.025
Apr 18 - Mar 19	0.07	0.07	0.01	0.030	0.030	0.031
Apr 19 - Mar 20	0.11	0.09	0.15	0.027	0.027	0.025
Apr 20 - Mar 21	0.12	-0.12	0.01	0.020	0.020	0.023

Table 4: Annual market performances of AT and other trading strategies.

As mentioned at the start of this subsection, there are many ways to refine the performance if our methodology is to be used for actual trading or full modelling of the exchange rate. In addition, a referee has pointed out that we could use traders' "buy" and "sell" signals to estimate their expectations. These can be used to augment the model's expectation of s_{t+1}, which differs from a trader's expectation. If such a refinement is sought in modelling exchange rates for trading, we suggest using a popular indicator from technical analysis called the Relative Strength Index (RSI), which traders often view as a signal to buy or sell.

Clearly, the AT methodology has brought forth a profitable trading strategy that dominates the traditional strategies on risk and return, and has also identified some areas of possible improvement. Our conclusion is that the continuing interest in the UIP (in spite of the forward-premium puzzle) was not misplaced: market forces have not, so far, uncovered the predictability in the exchange rate itself.

6 Concluding comments

Integration and co-integration have had a huge impact on the analysis of macroeconomic and aggregate financial data. They were a major first step in establishing methods to deal with the persistence of these variables. Here, we present an econometric method of analysis that is justified by the general-equilibrium solutions obtained in Abadir and Talmain (2002) which imply a specific type of persistence *and* cyclicality. We have applied it to solving an important puzzle in finance, but also to showing that substantial economic gains exist from knowing our process and that traders have so far been unable to capitalize on it. Our method has the potential to reveal new insights for other relations in macroeconomics and finance.

Appendix

This Appendix contains three parts. First, we present further analysis of the model in Section 2 then provide the proofs for the paper. Second, we discuss some features of our methodology. Third, we present the details of the method of one-step-ahead forecasting that we use.

Model and proofs

We can use the Cholesky decomposition to write the matrix $\boldsymbol{\Gamma}$ of (7) as $\boldsymbol{\Gamma} = \sigma^2 \boldsymbol{L}\boldsymbol{L}'$, where $\boldsymbol{L}^{-1}$ is the lower triangular matrix that removes the persistence from $\boldsymbol{u}$ and takes the form

$$\boldsymbol{L}^{-1} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ -\boldsymbol{\alpha}' & 1 \end{pmatrix}, \tag{A1}$$

with $\boldsymbol{\alpha} := (\alpha_{T-1}, \cdots, \alpha_2, \alpha_1)'$ and $\boldsymbol{A}$ a lower-triangular block of dimension $T-1$. Therefore, premultiplying (7) by $\boldsymbol{L}^{-1}$,

$$\boldsymbol{L}^{-1}\boldsymbol{y} = \boldsymbol{L}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ with } \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0},\ \sigma^2 \boldsymbol{I}_T). \tag{A2}$$

Notice that we reserve ε_t for well-behaved errors, and use u_t for errors with possible patterns such as long memory. In the special case of $\boldsymbol{\Gamma} = \sigma^2 \boldsymbol{I}_T$ (or $\boldsymbol{R} = \boldsymbol{I}_T$), the residuals $\boldsymbol{u}$ do not contain long memory anymore: we have $\boldsymbol{u} = \boldsymbol{\varepsilon}$ while long memory is allowed in the stationary $\boldsymbol{X}$.

The transformed residuals $\boldsymbol{\varepsilon}$ are now an uncorrelated sequence, and standard estimation procedures can be applied to the transformed model but, before doing so, we indicate that the estimates of $\boldsymbol{\varepsilon}$ can be used to determine the model's adequacy, as in standard setups. Such diagnostics include checking for leftover persistence, which can be due to any of: a simple omitted short-memory autocorrelation (which can be rectified by augmenting $\boldsymbol{X}$), an incorrect functional form for our ACF and its long-memory dynamics, and/or a spurious relation between $\boldsymbol{y}$ and $\boldsymbol{X}$. They also include a check for heteroskedasticity which, if untreated, would lead to finite-sample inefficient but still unbiased and consistent estimates. As in traditional approaches, this heteroskedasticity can be modelled and estimated directly (e.g., by adding a GARCH structure in $\boldsymbol{\varepsilon}$ and the corresponding likelihood) or indirectly by augmenting $\boldsymbol{X}$ with variables that were the source of the omitted heteroskedasticity.

The transformed data $\boldsymbol{L}^{-1}\boldsymbol{y}$ and $\boldsymbol{L}^{-1}\boldsymbol{X}$ in (A2) can be regressed by traditional methods. The Cholesky decomposition command is built-in as standard in all matrix-handling languages, such as Gauss and MATLAB. The Generalized Least Squares (GLS) estimators can be obtained by minimizing the criterion $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{\Gamma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ with respect to all parameters jointly. Alternatively, our ML estimators are obtained by maximizing the log-likelihood (apart from an additive constant)

$$-\frac{1}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{\Gamma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{A3}$$

where only the first term differs from the GLS criterion, and it has the beneficial effect of ensuring that the elements of the diagonal of $\boldsymbol{L}^{-1}$ are not too far from unity. This difference is responsible for another desirable property that the method of ML has, that it is invariant to reparameterizations of the model.

Proof of Proposition 1. Concentrating the normal log-likelihood with respect to

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} := (\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \text{ and } \widehat{\sigma}^2_{\boldsymbol{L}} := \frac{1}{T}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}})'\left(\boldsymbol{L}\boldsymbol{L}'\right)^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) \tag{A4}$$

by substituting $\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}, \widehat{\sigma}^2_{\boldsymbol{L}}$ for $\boldsymbol{\beta}, \sigma^2$ into (A3) and using $\boldsymbol{\Gamma} = \sigma^2\boldsymbol{L}\boldsymbol{L}' \propto \boldsymbol{R}$, we get

$$\begin{aligned} & -\log\left|\frac{1}{T}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}})'\left(\boldsymbol{L}\boldsymbol{L}'\right)^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}})\boldsymbol{L}\boldsymbol{L}'\right| - T \\ = \; & -\log\left|\frac{1}{T}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}})'\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}})\boldsymbol{R}\right| - T \\ = \; & -\log\left|(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}})'\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}})\boldsymbol{R}\right| + T\log(T) - T \end{aligned} \tag{A5}$$

since $|T^{-1}\boldsymbol{M}| = T^{-T}|\boldsymbol{M}|$ for any $T\times T$ matrix $\boldsymbol{M}$. Dropping the constant term $-T\left(1-\log(T)\right)$ yields (9), to be optimized with respect to the parameters of the ACF. ■

For the asymptotics of our ML procedure to apply more generally than for normally-distributed $\boldsymbol{\varepsilon}$, we assume the following regularity conditions with respect to the density from which the random (i.i.d.) sample $\boldsymbol{\varepsilon}$ is drawn. We denote by $\boldsymbol{\theta}$ the vector of parameters to be estimated:

Condition 1. The density is continuous in $\boldsymbol{\theta} \in \Theta$, and the true $\boldsymbol{\theta}$ (meaning the $\boldsymbol{\theta}$ that generated the data) belongs to the interior of the parameter space Θ.

Condition 2. In an open neighborhood of the true $\boldsymbol{\theta}$, the sample's log-likelihood $\ell(\boldsymbol{\theta})$ is continuously differentiable twice, and the expectations of $\ell(\boldsymbol{\theta})$ and its first two derivatives exist.

Condition 3. In an open neighborhood of the true $\boldsymbol{\theta}$, the sample's information matrix $\boldsymbol{\mathcal{I}}$ is positive definite, and it is an increasing function of T.

These conditions are standard; e.g., see Chapter 8 of Gouriéroux and Monfort (1995) or Chapter 12 of Abadir et al. (2018).

Proof of Proposition 2. The ACF of $\boldsymbol{u}$ satisfies $\rho_\tau = O(\tau^{-c})$. This is of the same order as an integrated process I(d) with $d = 1-2c < \frac{1}{2}$ when $d > 0$, and a short-memory process otherwise, hence satisfying Conditions 1–3 of Robinson and Hidalgo (1997) as discussed on

their p.83, and similarly for their pp.94-96 on the asymptotic equivalence of feasible GLS (the one with estimated $\boldsymbol{R}$) and infeasible GLS (the one with known $\boldsymbol{R}$). The latter also implies the asymptotic equivalence of the MLE $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{R}}}$ and infeasible GLS whose asymptotic normality is proven in Sections 2–3 of Robinson and Hidalgo (1997). ■

Proof of Proposition 3. Transform the data of each equation by $\widetilde{\boldsymbol{\Gamma}}$ (as in Section 3) and apply Proposition **2**. Then, the standard consistency of $\widehat{\boldsymbol{\Sigma}}$ implies that the asymptotic normality result applies to $\widetilde{\boldsymbol{\beta}}$. For $\widehat{\boldsymbol{\beta}}$, the same result follows by the optimization of the k terms in (16). ■

Remarks on the methodology

Before we comment on some aspects of our procedure, we indicate how it grew out of the treatment of models with autocorrelated errors, which are nested within our model. We take the simplest example

$$\begin{aligned} y_t &= \gamma x_t + u_t, \\ \text{with } u_t &= \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1, \quad \varepsilon_t \sim \text{IID}(0, \sigma^2). \end{aligned} \tag{A6}$$

To estimate this, taking into account the autocorrelation of u_t, the variables of the first equation (y_t and x_t) are transformed, then they are regressed by OLS to estimate the parameter γ of the relation. The vector $\boldsymbol{y} := (y_1, \ldots, y_T)'$ is transformed into

$$\boldsymbol{L}^{-1}\boldsymbol{y} \equiv \begin{pmatrix} \varphi & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\rho & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} \varphi y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{pmatrix}, \tag{A7}$$

where an estimate of ρ is substituted, and where φ is usually chosen as $\sqrt{1-\rho^2}$ to stabilize the variance of the transformed residuals. The lower triangular matrix $\boldsymbol{L}^{-1}$ that

premultiplies the vector of y_t-values arises from the Cholesky decomposition

$$\boldsymbol{R} = \begin{pmatrix} 1 & \rho & \rho^2 & \\ \rho & 1 & \rho & \ddots \\ \rho^2 & \rho & 1 & \ddots \\ & \ddots & \ddots & \ddots \end{pmatrix} = \left(1-\rho^2\right) \boldsymbol{L}\boldsymbol{L}'. \tag{A8}$$

Together with $\boldsymbol{\Gamma} = \sigma^2 \boldsymbol{L}\boldsymbol{L}'$, we see that the proportionality factor linking $\boldsymbol{R}$ to $\boldsymbol{\Gamma}$ is $\sigma^2/\left(1-\rho^2\right)$, the asymptotic variance of u_t.

If u_t were following an AR(p), then the lower triangular matrix $\boldsymbol{L}^{-1}$ in (A7) would contain $p+1$ nonzero diagonals, and the first p rows would have a normalization as was done for φ; e.g., see Chapter 5 of Amemiya (1985). When the variables have long memory, as is in our case, one needs a very large p to make this transformation. We overcame this problem by using our new ACF-based method in a parsimonious way. Using the matrix companion form, Abadir, Hadri, and Tzavalis (1999) showed that long lags have a similar effect to adding dimensions to a VAR, which would increases the bias and variance of the estimators. Finding a parsimonious solution avoids these types of problems.

We make the following remarks on the requirements and/or features of $\boldsymbol{R}$ and the corresponding $\boldsymbol{L}$ in our procedure:

Remark A1. In estimating the parameters of the ACF, one needs to restrict their values so that the estimated $\widehat{\boldsymbol{R}}$ is positive definite, since this is true (by definition) for $\boldsymbol{R}$. There is no explicit formula for this restriction, because there is no explicit solution for the roots of polynomials of order greater than 4. Nevertheless, it is straightforward to implement the restriction numerically either by skipping values that do not satisfy the restriction, or by imposing a Lagrangian penalty in the objective (e.g., log-likelihood) function.

Remark A2. The lower triangularity of $\boldsymbol{L}^{-1}$ ensures that each element of the transformed $\boldsymbol{y}$ is constructed only from past and current (but no future) values of y_t; e.g., see (A7). The same comment applies to $\boldsymbol{X}$.

Remark A3. The elements in the last row of $\boldsymbol{L}^{-1}$ have an interpretation as the coefficients of an AR($T-1$) representation for the last transformed data point, which is why we stated them explicitly in (A1). Note that *any* non-explosive process, whether nonlinear and/or nonstationary, can be represented as an invertible MA having time-varying coefficients, which explains the time-varying AR representations implied by the rows of $\boldsymbol{L}^{-1}$. This is known in time series as Cramér's decomposition, a generalization of Wold's decomposition; see Granger and Newbold (1986) and the widespread applications in McCabe, Martin, and Tremayne (2005). It is why the nonlinear process arising from Abadir and Talmain (2002) can be estimated by our linear representation (with square-summable coefficients). We also refer to Baillie, Diebold, Kapetanios, and Kim (2022) and to sieve regressions as an alternative way to flexibly model unspecified dynamics (instead of our ACF's implied specific ARs in $\boldsymbol{L}^{-1}$) in the transformed (A2). Here, we allow for long-memory and time-varying AR representations for each data point, the last one being an AR($T-1$), as implied by (A1). The coefficients of these time-varying ARs follow directly from only the 4 parameters in the long-memory ACF.

Remark A4. A well-known feature of the transformed model (A2) is that the constant, once transformed by $\boldsymbol{L}^{-1}$, is not a constant vector anymore; e.g., use $\boldsymbol{\imath} := (1,\ldots,1)'$ instead of $\boldsymbol{y}$ in (A7) and compare the first element to the remaining $T-1$. In our estimations, it is therefore assumed that the data ($\boldsymbol{y}$ and $\boldsymbol{X}$) have been de-meaned before being transformed. This is because the procedure is based on transforming vectors, say $\boldsymbol{z}$, which are centered around $\boldsymbol{0}$: from $\boldsymbol{z} \sim \mathrm{D}(\boldsymbol{0}, \boldsymbol{L}\boldsymbol{L}')$ into $\boldsymbol{L}^{-1}\boldsymbol{z} \sim \mathrm{D}(\boldsymbol{0}, \boldsymbol{I}_T)$. Having a nonzero sample mean in $\boldsymbol{z}$ would have introduced a common term like $\boldsymbol{L}^{-1}\boldsymbol{\imath}$ in all these transformed variables, which may dominate these series and produce some seemingly common factor that causes multicollinearity and other unnecessary numerical instabilities. If a constant is required in the regression, it should be transformed separately then added to the regression for transformed variables. Numerical instabilities apart, the theorem of Frisch and Waugh

(1933) proves that the resulting point estimates would be identical with or without removing the mean.

Forecasting

We conclude this Appendix with our procedure for 1-step-ahead forecasting. Starting from (A2), we rewrite it as

$$\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{A9}$$

where $\widetilde{\boldsymbol{y}} := \boldsymbol{L}^{-1}\boldsymbol{y}$ and $\widetilde{\boldsymbol{X}} := \boldsymbol{L}^{-1}\boldsymbol{X}$. Since the residuals are now $\boldsymbol{\varepsilon}$, traditional methods of forecasting can be used on this reformulated model, once the variables are transformed into $\widetilde{\boldsymbol{y}}$ and $\widetilde{\boldsymbol{X}}$. Furthermore, if desired it is now possible to re-estimate $\boldsymbol{\beta}$ recursively out-of-sample using this equation, although we did not do so because of the stability of our estimates of $\boldsymbol{\beta}$. To make all this operational, we now give the details of the transformation of the variables by our estimator of $\boldsymbol{L}$.

We denote the estimation period by $t = 1, \ldots, T_1 < T$, then 1-step-ahead forecasts $\widehat{y}_t$ are needed for y_t over $t = T_1 + 1, \ \ldots, \ T$. Because of the dynamic nature of the model, the transformation of y_t into $\widetilde{y}_t$ will require past values of y_t and we use data starting in period $T_0 < T_1$, with $T_0 \ll T_1$ in long memory models. Empirically, a good choice for the initial number of lags $N_{\text{ini}} := T_1 - T_0 + 1$ appears to be around $3T_\omega$, where $T_\omega := 2\pi/\omega$ is the period of the cycle associated with ω.

Let $N := T - T_0 + 1$. The next step is to fill an $N \times N$ matrix $\boldsymbol{R}$ by using the ACF parameters estimated from the first subsample, and we denote the corresponding estimate of $\boldsymbol{L}$ by $\widehat{\boldsymbol{L}}$ and use it to calculate

$$\widetilde{\boldsymbol{y}} := \widehat{\boldsymbol{L}}^{-1}\boldsymbol{y} \qquad \text{and} \qquad \widetilde{\boldsymbol{X}} := \widehat{\boldsymbol{L}}^{-1}\boldsymbol{X} \tag{A10}$$

for $\boldsymbol{y} = (y_{T_0}, \ldots, y_T)'$ and similarly for $\boldsymbol{X}$. Note the following:

1. To avoid complicating the notation, we wrote $\widetilde{\boldsymbol{y}}$ instead of $\widehat{\widetilde{\boldsymbol{y}}}$ since there are no cases where the unknown $\boldsymbol{L}^{-1}\boldsymbol{y}$ is used simultaneously with the known $\widehat{\boldsymbol{L}}^{-1}\boldsymbol{y}$. We will reserve $\widehat{\widetilde{\boldsymbol{y}}}$ for the forecast of $\widetilde{\boldsymbol{y}}$.

2. We transform all of the last N observations simultaneously in (A10).

3. As shown earlier in this Appendix, $\boldsymbol{L}^{-1}$ is lower-triangular and hence uses only past values to transform any y_t into $\widetilde{y}_t$, and similarly here for $\widehat{\boldsymbol{L}}^{-1}$.

The forecast for $\widetilde{y}_{t+1}$ is then obtained by setting ε_{t+1} to its expected value of zero, yielding

$$\widehat{\widetilde{\boldsymbol{y}}} = \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} \tag{A11}$$

as the forecast. It is the optimal forecast, since we condition on the first subsample (the observed past) and its estimation of the ACF parameters which gave us the $\widehat{\boldsymbol{L}}$ used in (A10).

This is a forecast for the transformed variables $\widetilde{y}_{t+1}$, but we need a forecast for the original variables y_{t+1}, and this is when the forecast becomes a one-step-ahead forecast for y_{t+1} because we will be multiplying the relevant vector of $\widetilde{y}$'s and $\widehat{\widetilde{y}}$'s by $\widehat{\boldsymbol{L}}$. For $t = T_1, \ldots, T-1$, we have a choice of using either

$$\begin{pmatrix} \widetilde{y}_{T_0+1} \\ \vdots \\ \widetilde{y}_t \\ \widehat{\widetilde{y}}_{t+1} \\ 0 \\ \mathbf{0}_{T-t-1} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \widetilde{y}_{T_0} \\ \widetilde{y}_{T_0+1} \\ \vdots \\ \widetilde{y}_t \\ \widehat{\widetilde{y}}_{t+1} \\ \mathbf{0}_{T-t-1} \end{pmatrix} \tag{A12}$$

to premultiply it by the t-th row or the $(t+1)$-th row of $\widehat{\boldsymbol{L}}$, respectively, thus yielding two alternative forecasts $\widehat{y}_{t+1}$. Note the following:

1. At time t, we know $\widetilde{y}_{t-j}$ for $j \geq 0$, which are therefore used in the early rows of (A12) instead of the past forecasted $\widehat{\widetilde{y}}_{t-j}$ ($\neq \widetilde{y}_{t-j}$ with probability 1).

2. The first of the two possible vectors in (A12) drops the initial $\widetilde{y}_{T_0}$ to accommodate $\widehat{\widetilde{y}}_{t+1}$ as the last nonzero element and to stay on the same t-th row as when the forecast was constructed using the $y_{T_0}, \ldots, y_t$ (which were transformed by the t-th row in $\widehat{\boldsymbol{L}}^{-1}\boldsymbol{y}$),

but the second choice enlarges the nonzero elements of the vector to include the $\widehat{\widetilde{y}}_{t+1}$ while keeping the initial $\widetilde{y}_{T_0}$. The former uses a lag polynomial implied by the t-th row of $\widehat{\boldsymbol{L}}^{-1}$ and its corresponding inverse in the t-th row of $\widehat{\boldsymbol{L}}$, but drops the initial $\widetilde{y}_{T_0}$ in return. We found that the two methods gave essentially the same answer in our application, but this need not be so in general.

Finally, we note that in the special case of our application, we could have done the forecasting through $\widehat{s}_{t+1} = \widehat{\theta} s_t + \widehat{u}_{t+1}$, since we do not re-estimate $\widehat{\theta} = 0.88$. In this case, we would need a forecast $\widehat{u}_{t+1}$. It could be based on the general method in (A12) or on $\boldsymbol{\alpha}$ in (A1) which takes the longest lag polynomial although truncation is recommended in dealing with long-memory processes.

References

Abadir, Karim M., Giovanni Caggiano, and Gabriel Talmain, 2013, Nelson-Plosser revisited: the ACF approach, *Journal of Econometrics* 175, 22-34.

Abadir, Karim M., Walter Distaso, and Liudas Giraitis, 2009, Two estimators of the long-run variance: beyond short memory, *Journal of Econometrics* 150, 56-70.

Abadir, Karim M., Kaddour Hadri, and Elias Tzavalis, 1999, The influence of VAR dimensions on estimator biases, *Econometrica* 67, 163-181.

Abadir, Karim M., Risto D. H. Heijmans, and Jan R. Magnus, 2018, *Statistics* (Cambridge University Press, Cambridge).

Abadir, Karim M., and Gabriel Talmain, 2002, Aggregation, persistence and volatility in a macro model, *Review of Economic Studies* 69, 749-779.

Abadir, Karim M., and A.M. Robert Taylor, 1999, On the definitions of (co-)integration, *Journal of Time Series Analysis* 20, 129-137.

Amemiya, Takeshi, 1985, *Advanced Econometrics* (Harvard University Press, Cambridge).

Apte, Prakash, Piet Sercu, and Raman Uppal, 2004, The exchange rate and purchasing power parity: extending the theory and tests, *Journal of International Money and Finance* 23, 553-571.

Backus, David K., Allan W. Gregory, and Chris I. Telmer, 1993, Accounting for forward rates in markets for foreign currency, *Journal of Finance* 48, 1887-1908.

Baillie, Richard T., and Tim Bollerslev, 1994, Cointegration, fractional cointegration, and exchange rate dynamics, *Journal of Finance* 49, 737-745.

Baillie, Richard T., and Tim Bollerslev, 2000, The forward premium anomaly is not as bad as you think, *Journal of International Money and Finance* 19, 471-488.

Baillie, Richard T., Francis X. Diebold, George Kapetanios, and Kun Ho Kim, 2022, Robust inference in time series regression, PIER Working Paper 22-012.

Bansal, Ravi, and Ivan Shaliastovich, 2013, A long-run risks explanation of predictability puzzles in bond and currency markets, *Review of Financial Studies* 26, 1-33.

Bekaert, Geert, 1996, The time variation of risk and return in foreign exchange markets: a general equilibrium perspective, *Review of Financial Studies* 9, 427-470.

Beran, Jan, 1992, Statistical methods for data with long-range dependence, *Statistical Science* 7, 404-427 (with discussion).

Carvalho, Carlos, and Fernanda Nechio, 2011, Aggregation and the PPP Puzzle in a Sticky-Price Model, *American Economic Review* 101, 2391-2424.

Chambers, Marcus J., 1998, Long memory and aggregation in macroeconomic time series, *International Economic Review* 39, 1053-1072.

Chauvet, Marcelle, and Simon Potter, 2013, Forecasting output, *Handbook of Economic Forecasting* 2, 141-194.

Chevillon, Guillaume, and Sophocles Mavroeidis, 2017, Learning can generate long memory, *Journal of Econometrics* 198, 1-9.

Chevillon, Guillaume, Alain Hecq, and Sébastien Laurent, 2015, Long memory through marginalization of large systems and hidden cross-section dependence, SSRN paper (https://ssrn.com/abstract=2612870).

Della Corte, Pasquale, Lucio Sarno, and Ilias Tsiakas, 2009, An economic evaluation of empirical exchange rate models, *Review of Financial Studies* 22, 3491-3530.

Diebold, Francis X., Javier Gardeazabal, and Kamil Yilmaz, 1994, On cointegration and exchange rate dynamics, *Journal of Finance* 49, 727-735.

Diebold, Francis X., Steven Husted, and Mark Rush, 1991, Real exchange rates under the gold standard, *Journal of Political Economy* 99 1252-1271.

Engel, Charles, 1996, The forward discount anomaly and the risk premium: a survey of recent evidence, *Journal of Empirical Finance* 3, 123-192.

Engel, Charles, 2016, Exchange rates, interest rates, and the risk premium, *American Economic Review* 106, 436-174.

Engel, Charles, and James D. Hamilton, 1990, Long swings in the dollar: are they in the data and do markets know it? *American Economic Review* 80, 689-713.

Engel, Charles, and Kenneth D. West, 2005, Exchange rates and fundamentals, *Journal of Political Economy* 113, 485-517.

Fisher, Irving, 1930, *The Theory of Interest* (Macmillan, New York).

Frisch, Ragnar, and Frederick V. Waugh, 1933, Partial time regressions as compared with individual trends, *Econometrica* 1, 387-401.

Froot, Kenneth A., and Richard H. Thaler, 1990, Foreign exchange, *Journal of Economic Perspectives* 4, 179-192.

Gil-Alana, Luis A., 2001, Testing stochastic cycles in macroeconomic time series, *Journal of Time Series Analysis* 22, 411-430.

Gil-Alana, Luis A., and Tommaso Trani, 2019, The cyclical structure of the UK inflation rate: 1210–2016, *Economics Letters* 181, 182-185.

Giraitis, L., J. Hidalgo, and P.M. Robinson, 2001, Gaussian estimation of parametric spectral density with unknown pole, *Annals of Statistics* 29, 987-1023.

Gouriéroux, Christian, and Alain Monfort, 1995, *Statistics and Econometric Models, volume 1* (Cambridge University Press, Cambridge).

Granger, Clive W.J., 1980, Long memory relationships and the aggregation of dynamic models, *Journal of Econometrics* 14, 227-238.

Granger, Clive W.J., and Paul Newbold, 1986, *Forecasting Economic Time Series* (Academic Press, San Diego).

Guidolin, Massimo, Stuart Hyde, David McMillan, and Sadayuki Ono, 2009, Non-linear predictability in stock and bond returns: When and where is it exploitable? *International Journal of Forecasting* 25, 373-399.

Hassan, Tarek A., and Rui C. Mano, 2019, Forward and spot exchange rates in a multi-currency world, *The Quarterly Journal of Economics* 134, 397-450.

Hidalgo, Javier, 2005, Semiparametric estimation for stationary processes whose spectra have an unknown pole, *Annals of Statistics* 33, 1843-1889.

Hong, Yongmiao, and Yoon Jin Lee, 2005, Generalized spectral tests for conditional mean models in time series with conditional heteroscedasticity of unknown form, *Review of Economic Studies* 72, 499-541.

Jentsch, Carsten, and Dimitris N. Politis, 2015, Covariance matrix estimation and linear process bootstrap for multivariate time series of possibly increasing dimension, *Annals of Statistics* 43, 1117-1140.

Kostakis, Alexandros, Tassos Magdalinos, and Michalis P. Stamatogiannis, 2015, Robust econometric inference for stock return predictability, *Review of Financial Studies* 28, 1506-1553.

Kruse, Robinson, 2011, A new unit root test against ESTAR based on a class of modified statistics, *Statistical Papers* 52, 71-85.

Lewis, Karen K., 1995, Puzzles in international financial markets, in G. Grossman and K. Rogoff, eds.: *Handbook of International Economics*, vol. 3 (North-Holland, Amsterdam).

Maddala, G.S., and A.S. Rao, 1973, Tests for serial correlation in regression models with lagged dependent variables and serially correlated errors, *Econometrica* 47, 761-774.

Maynard, Alex, and Peter C.B. Phillips, 2001, Rethinking an old empirical puzzle: econometric evidence on the forward discount anomaly, *Journal of Applied Economics* 16, 671-708.

McCabe, B. P. M., G. M. Martin, and A. R. Tremayne, 2005, Assessing persistence in discrete nonstationary time-series models, *Journal of Time Series Analysis* 26, 305-317.

McMurry, Timothy L., and Dimitris N. Politis, 2010, Banded and tapered estimates for autocovariance matrices and the linear process bootstrap, *Journal of Time Series Analysis* 31, 471-482.

Okunev, John, and Derek White, 2003, Do momentum-based strategies still work in foreign currency markets? *Journal of Financial and Quantitative Analysis* 38, 425-447.

Pierce, Donald A., 1982, The asymptotic effect of substituting estimators for parameters in certain types of statistics, *Annals of Statistics* 10, 475-478.

Robinson, P.M., 1978, Statistical inference for a random coefficient autoregressive model, *Scandinavian Journal of Statistics* 5, 163-168.

Robinson, P.M., 1994, Time series with strong dependence, in C.A. Sims, ed.: *Advances in Econometrics: sixth world congress*, vol.1 (Cambridge University Press, Cambridge).

Robinson, P.M., and F.J. Hidalgo, 1997, Time series regression with long-range dependence, *Annals of Statistics* 25, 77-104.

Sarno, Lucio, Giorgio Valente, and Hyginus Leon, 2006, Nonlinearity in deviations from uncovered interest parity: an explanation of the forward bias puzzle, *Review of Finance* 10, 443-482.

Talmain, Gabriel, 2018, Two-country model and foreign exchange dynamics, SSRN paper (https://ssrn.com/abstract=3140312).

Verdelhan, Adrien, 2010, A habit-based explanation of the exchange rate risk premium, *Journal of Finance* 65, 123-146.

Zellner, Arnold, 1962, An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American Statistical Association* 57, 348-368.

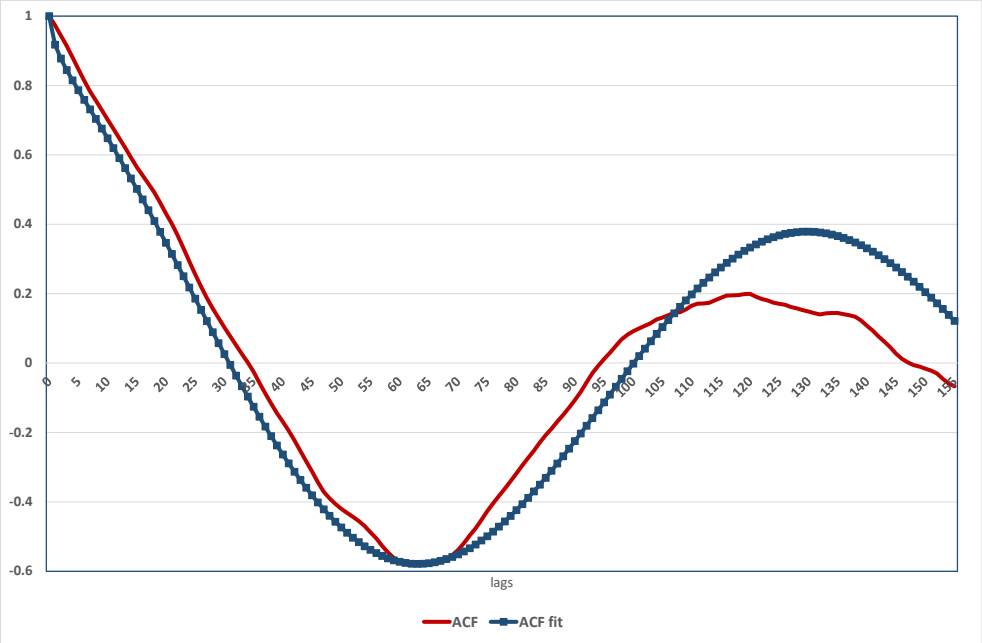

Figure 1. ACF of the logarithm of the GBP-USD exchange rate and its fit by our functional form, August 1976 to March 2015.

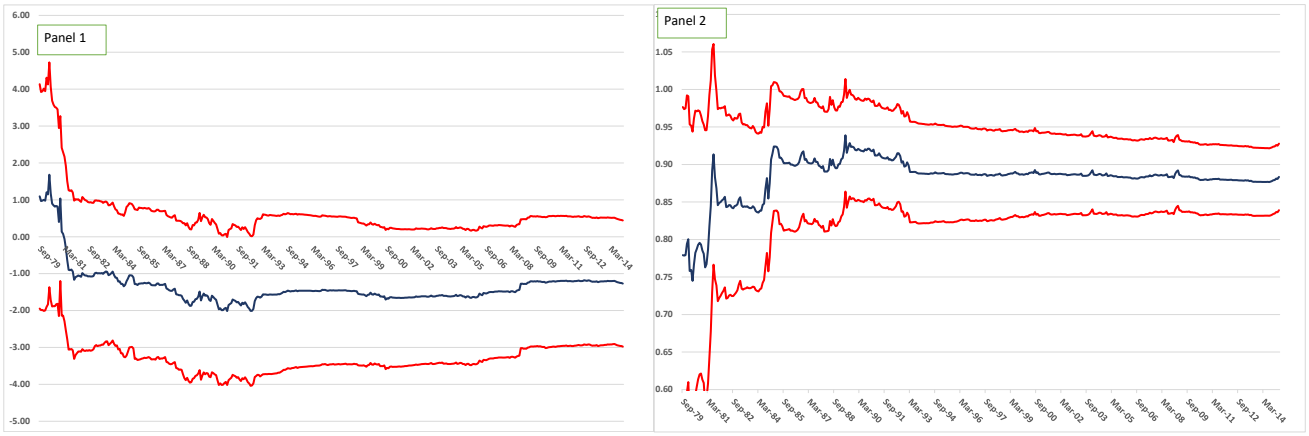

Figure 2a. Recursive parameter estimates as the sample size is increased, with approximate ±2 SE bands, for the AT model (25). Panels 1–2 represent the parameter estimates corresponding to $f_t - s_t$, and s_t, respectively.

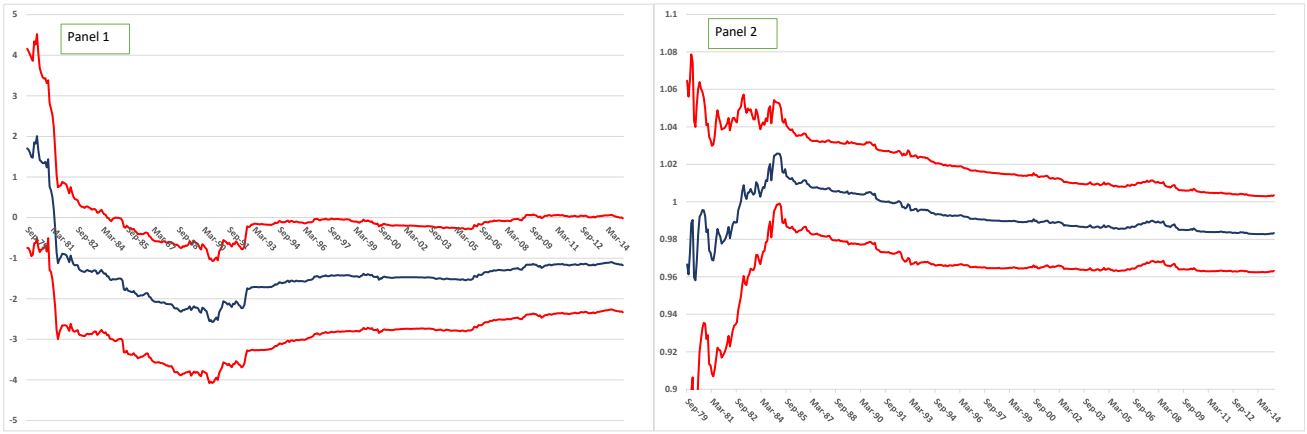

Figure 2b. Recursive parameter estimates as the sample size is increased, with approximate ±2 SE bands, for the NoD model (26). Panels 1–2 represent the parameter estimates corresponding to $f_t - s_t$, and s_t, respectively.

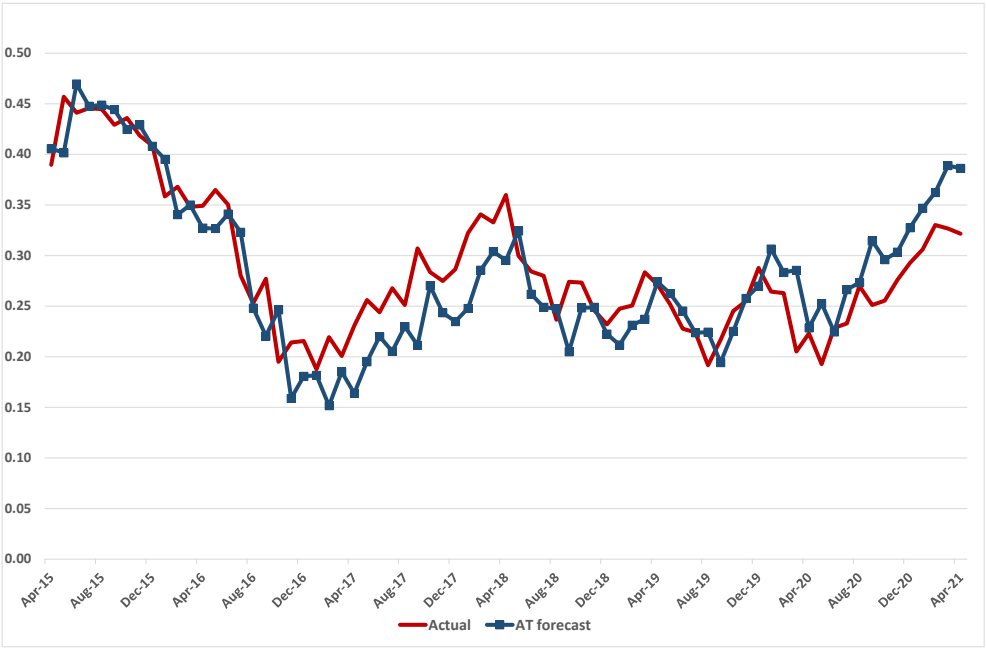

Figure 3: Spot rate, actual vs AT out-of-sample forecast.

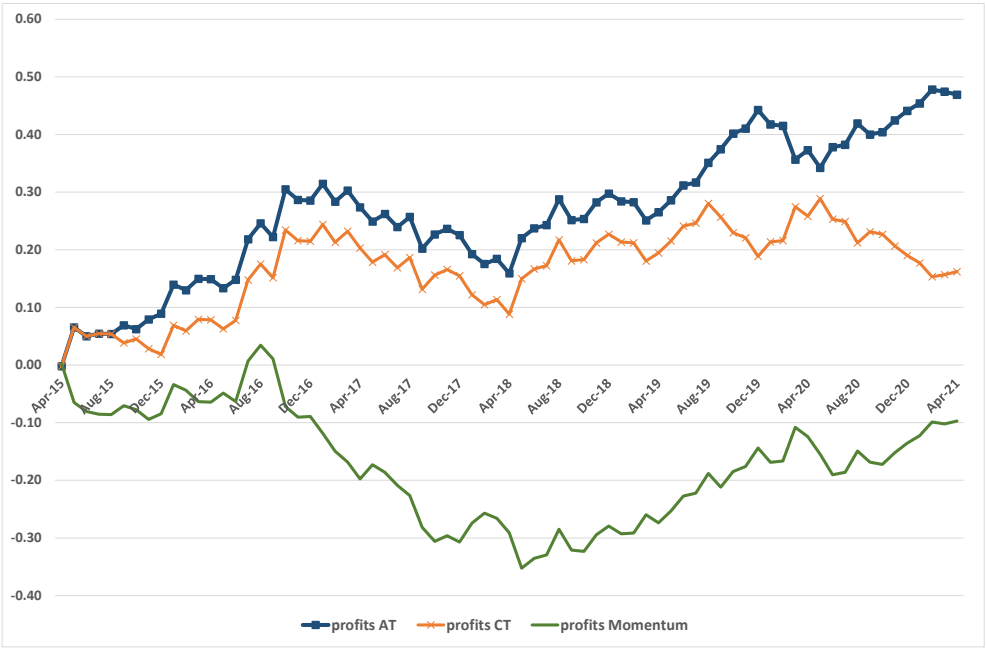

Figure 4. Out-of-sample cumulative profits of FX strategies: AT, Carry Trade (CT), and Momentum.