Combining Long Memory and Level Shifts in Modeling and Forecasting the Volatility of Asset Returns:
Supplementary Appendix

Rasmus T. Varneskov∗ Pierre Perron†
* Aarhus University and CREATESE † Boston University

January 30, 2017

Abstract

This online appendix supplements the main text, Varneskov & Perron (2017), with further descriptions of theory, in particular for the RLS-ARFIMA model and different semi-parametric time series tools. Moreover, it provides implementation details, additional empirical results as well as the proofs of Propositions 1 and 2. Its respective sections are referred to throughout the main text and may be read independently without going through the whole supplementary material. They are, however, based on notation and definitions from the main text.

∗Department of Economics and Business, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V., Denmark. Email: rvarneskov@creates.au.dk.
†Department of Economics, Boston University, 270 Bay State Rd., Boston, MA, 02215. Email: perron@bu.edu.
1 Introduction

This online appendix supplements the main text, Varneskov & Perron (2017), with additional descriptions of theory and tools, empirical results as well as proofs. In particular, Section 2 describes the relation between discrete and continuous time volatility measures, and it presents the generalized flat-top realized kernel estimator of Varneskov (2016a, 2016b), which is used to construct the high-frequency (HF) volatility measures for the empirical analysis. Section 3 relates the discrete time RLS-ARFIMA model to a recently proposed continuous time stochastic volatility model. Section 4 provides a theoretical discussion of the autocorrelation function for an RLS-ARFIMA process, it describes the log-periodogram and local Whittle estimators as well as the tests by Perron & Qu (2010) and Qu (2011), and it gives some empirical evidence. Section 5 details the maximum likelihood estimation procedure. Section 6 briefly reviews implementation of the HAR and GARCH models, and Section 7 lays out the Model Confidence Set (MCS) testing procedure of Hansen, Lunde & Nason (2011). Finally, proofs of Varneskov & Perron (2017, Propositions 1 and 2) are provided in Sections 8 and 9, respectively. All sections are self-contained in the sense that they may be read without going through the entire supplementary appendix. They are, however, based on notation and definitions from the main text.

2 Volatility Measurement

This section briefly discusses and relates volatility measurement for discrete and continuous time models, and it presents the generalized flat-top realized kernel estimator of Varneskov (2016a, 2016b), which is used to construct the HF volatility measures for the empirical analysis.

2.1 Discrete- and Continuous-time SV Frameworks

First, let $p_t = \ln(R_t) \in \mathbb{R}$ and $r_t = p_t - p_{t-1} \in \mathbb{R}$ for $t = 1, \ldots, T$ denote a sequence of univariate daily logarithmic asset prices and returns, respectively, which we stipulate to obey a discrete-time innovation model of the form

$$ r_t = V_t z_t, \quad z_t \sim \text{i.i.d.} N(0, 1). \quad (1) $$

While simple, this innovation model is consistent with asset pricing theory, which contends that the current asset price reflects the discounted value of expected future cash flows. The time-varying volatility, $V_t$, governs the scale of the log-return distribution to capture stylized empirical facts about asset returns such as, e.g., volatility clustering. Depending on the specification of the latent (log-)volatility process, this model embodies the classical discrete time (G)ARCH, and stochastic volatility (SV) frameworks, c.f. the reviews in, e.g., Shephard (2005), Teräsvirta (2009) and Andersen & Benzoni (2012).

The recent availability of high-frequency financial data, however, has produced dramatic changes to the way daily variability is measured and, subsequently, modeled. Since asset prices are recorded at every intra-daily trade, (1) inadequately describes the dynamic evolution of the return and/or volatility process during a trading day. Hence, again with origin in standard asset pricing theory, the logarithmic
asset price is, instead, assumed to follow a continuous-time Brownian semimartingale with stochastic volatility and, possibly, jumps defined on a filtered probability space \((\Omega, \mathcal{F}, (\mathcal{F}_s)_{s \geq 0}, \mathbb{P})\). In particular, denote the logarithmic price process on a given trading day, \(t\), by \(p(s_t)\) for \(s_t \in [t - 1, t]\) and define the continuously compounded log-return over the period \([s_t - h, s_t]\) for \(t - 1 \leq t - 1 + h \leq s_t \leq t\) as
\[
 r(s_t, h) = p(s_t) - p(s_t - h) = \int_{s_t - h}^{s_t} a(u)du + \int_{s_t - h}^{s_t} V(u)du + \int_{s_t - h}^{s_t} \kappa(u)du(u) \tag{2}
\]
where the drift, \(a(u)\), and spot volatility, \(V(u)\), are locally bounded and càdlàg, \(W(u)\) is a standard Brownian motion and, finally, \(q(u)\) is a finite activity counting process where \(dq(u) = 1\) corresponds to a jump at time \(u\) with magnitude \(\kappa(u)\). In this setting, we are specifically interested in modeling the quadratic variation of (2), which is defined as
\[
 V_t^2 = \int_{t-1}^{t} V(u)^2du + \sum_{t-1 \leq s \leq t} \kappa(s)^2, \tag{3}
\]
or, more precisely, in modeling its square-root transformation, \(V_t\). The continuous-time return and return variation in (2) and (3), respectively, are intimately linked to the model in (1). To see this, assume for simplicity that \(r(s_t, h)\) exhibits neither drift nor jumps, i.e. \(a(u) = 0\), \(dq(u) = 0\) \(\forall u \in [t - 1, t]\), then we may, under appropriate assumptions on (2), write a sequence of continuously compounded daily log-returns as
\[
 r(t, 1)|\mathcal{F}_t \sim N \left(0, V_t^2\right), \quad t = 1, \ldots, T, \tag{4}
\]
thus with a similar innovation representation where (the square-root of) quadratic variation replaces the daily volatility measure, \(V_t\). In fact, quadratic variation is an ideal measure of asset return variability since it captures the entire path of the volatility process.\(^1\) We stress that both the conditional distribution in (4) and quadratic variation are unobserved since they are contingent on the ex-post realization of \(V(u)\) over a given interval \([t - 1, t]\), which can only be obtained with a continuous and frictionless price record that does not exist in practice. Both quantities may, however, be precisely estimated using discretely sampled high-frequency asset returns if the applied estimator is able to account for the array of market frictions that are inherent to observed (log-)prices at higher frequencies.

### 2.2 The Flat-top Realized Kernel Approach

Suppose that for a given asset on a given trading day \(t\) we observe \(n_t + 1\) discrete intra-daily log-prices, \(z_{t_i} = p_{t_i} + u_{t_i}, \ i = 0, \ldots, n_t\), which consist of a signal, \(p_{t_i}\), for example the model in (2)-(4), and a market microstructure (MMS) noise component, \(u_{t_i}\), that summarizes a diverse array of market imperfections, e.g., bid-ask bounce effects and asymmetric information among market participants.\(^2\) In this setting, we cannot apply standard estimators of quadratic variation such as realized variance, since the presence of

\(^1\)Andersen, Bollerslev & Diebold (2008) and Barndorff-Nielsen & Shephard (2007) review the use of quadratic variation in the economics of financial risk and provide extensive references to literature.
\(^2\)The observations are assumed to be scattered on a sampling grid \(t - 1 \leq t_0 < t_1 \cdots < t_{n_t} \leq t\), which may be random.
MMS noise renders them inconsistent as the sampling interval progressively shrinks. Hence, to obtain precise and robust estimates of quadratic variation, we first let \( \Delta = 1 - L \) be the usual differencing operator, \( S^+_h = \max(h, 0) \), \( S^-_h = \min(h, 0) \) and define the realized autocovariances of the HF log-returns as

\[
\Gamma_h(z_t) = \sum_{i=1+S^+_h}^{n_t+S^+_h} \Delta z_t \Delta z_{t-i-h}, \quad \forall h = -(n_t - 1), \ldots, -1, 0, 1, \ldots, n_t - 1.
\]

These form the building blocks for a general class of realized kernel estimators,

\[
RK(z_t) = \Gamma_0(z_t) + \sum_{h=1}^{n_t-1} k \left( \frac{h}{H_t} \right) \{ \Gamma_h(z_t) + \Gamma_{-h}(z_t) \},
\]

where \( k(h/H_t) \) is a non-stochastic weight function and \( H_t = \alpha n_t^{1/2}, \alpha > 0, \) is a bandwidth. The properties of realized kernel estimators, however, depend crucially on the shape of the kernel function, particularly around the origin. Varneskov (2016a, 2016b) show that by selecting a flat-top kernel,

\[
k(x_t) = 1\{|x_t| \leq c_t\} + \lambda(|x_t| - c_t)1\{|x_t| > c_t\},
\]

for some shrinking function of the bandwidth \( c_t = H_t^{-\gamma_c}, \gamma_c \in (0, 1) \), as well as a second-order smooth kernel function \( \lambda(\cdot) \), which satisfies mild regularity conditions, the resulting generalized flat-top realized kernel estimators of quadratic variation have optimal asymptotic properties such as consistency, asymptotic unbiasedness and mixed Gaussianity at the (optimal) rate of convergence, \( n_t^{1/4} \), under mild assumptions on the MMS noise and provided that the shrinkage parameter, \( \gamma_c \), is chosen sufficiently small. If optimally designed, the estimators are also efficient in a Cramér-Rao sense.

In addition to their asymptotic properties, simulations illustrate that the flat-top realized kernel estimators have a desirable combination of robustness and efficiency in finite samples, even for data sampled as sparsely as, e.g., every minute. Following implementation recommendations, we select the weight function \( \lambda(\cdot) \) to be the Parzen kernel, \( \gamma_c = 3/5 \) as the flat-top shrinkage, and the bandwidth parameter \( a \) according to Varneskov (2016a, Section 5.2).

### 3 RLS-ARFIMA and Continuous Time SV Models

Our objective in the main text is to model and forecast various daily volatility series using the reduced form volatility model in Varneskov & Perron (2017, Equation (4)). However, there are, in fact, striking parallels between our discrete time model and the continuous time stochastic volatility (SV) literature, which deserve a few comments. To clarify these similarities, suppose that \( h_t \) follows an AR(1) process with persistence parameter \( \phi \in [0, 1) \) and that there are no measurement errors in \( y_t \), that is, \( u_t = 0 \) for all \( t \), then we may write

\[
\Delta y_t = (\phi - 1)(y_{t-1} - a - v_{t-1}) + \epsilon_t + \Delta v_t
\]
where $\Delta v_t$ are the innovations to the random level shift process, $\epsilon_t$ are the Gaussian innovations, and we have $(\phi - 1) < 0$. The first term may be interpreted as a "mean-reverting" drift where, however, the mean $a + v_{t-1}$ is determined by infrequent breaks. The second component, $\epsilon_t + \Delta v_t$, captures Gaussian and non-Gaussian innovations to the volatility path, with $\Delta v_t$ being a simple Lévy process. As the increments to the random level shift process are independent, so are $v_{t-1}$ and $\Delta v_t$. Indeed, it is the specification of volatility innovations, which has been analyzed in much of the recent work on continuous time SV models.\textsuperscript{3} Hence, to illustrate how our discrete time model relates to this literature, let us write the log-volatility model of, e.g., Todorov et al. (2014), using their notation, as

$$d\ln V_t = \mu + dv_t, \quad dv_t = -\kappa dt + dL_t$$

where $\mu$ is the unconditional mean, $\kappa > 0$ captures the speed of mean reversion, and $L_t$ is a general Lévy specification. Similar to (5), this model decomposes into a drift and a Lévy component, which is Gaussian if its activity index is 2. The activity index, thus, quantifies the relative importance of Gaussian and non-Gaussian innovations. In our discrete time model (5), on the other hand, this is conveyed through the relative magnitudes of $\sigma_\epsilon$ and $(\sigma_n, \gamma)$. This clearly shows that our reduced form volatility model is related to contemporaneous continuous time volatility models. However, instead of providing a detailed analysis of the specific form of Lévy innovations, as in Todorov et al. (2014), we are concerned with the mean of the process, which may be time-varying, the speed of mean reversion, which may be determined by a fractional filter, and their implications for forecasting.

4 A Gauge of Volatility Dynamics

As motivation, and of separate interest to the reader, this section considers the implications of the model in Varneskov & Perron (2017, Equations (2)-(3)) on the autocovariance and autocorrelation functions as well as on the periodogram of the series, and we gauge these features empirically. Finally, the testing procedures by Perron & Qu (2010) and Qu (2011) are described.

4.1 The Autocovariance and Autocorrelation Functions

The asymptotic properties of the autocovariance and autocorrelation functions for a short memory random level shift model, derived in Perron & Qu (2010), may readily be generalized to accommodate genuine long memory and measurement errors in the volatility series. A crucial ingredient for this generalization is the functional central limit theorem for the cumulative random level shift process, $v_t$, considered by Georgiev (2002) and Leipus & Viano (2003). They derived the following weak convergence result under the Skorohod topology, denoted by $\Rightarrow$.\textsuperscript{4}

\textsuperscript{3}See, for example, the contributions by Aít-Sahalia & Jacod (2009), Todorov & Tauchen (2011), Andersen, Bondarenko, Todorov & Tauchen (2015), Todorov, Tauchen & Grynkiv (2014), and many references therein.

\textsuperscript{4}See also the discussion in Perron & Qu (2007) for details on the properties of the random level shift process.
Lemma A1 (Georgiev (2002) and Leipus & Viano (2003)). Let the random level shift component, $v_t$, be defined as in Varneskov & Perron (2017, Section 2) with $\gamma > 0$ fixed, then

$$v_{Ts} = \sum_{j=1}^{[Ts]} \delta_{T,j} \Rightarrow J(s), \quad \text{where} \quad J(s) = \sum_{j=0}^{N(s)} \eta_j,$$

with $N(s)$ being a Poisson process with jump intensity $\gamma$ that is independent of $\eta_j$ for all $j$.

Recall that the autocovariance function of $h_t$, the latent ARFIMA component in the signal-plus-noise model of Varneskov & Perron (2017, Equations (2)-(3)), is denoted by $R_h(\tau)$, and let the sample autocovariance function for the observable log-volatility process, $y_t$, be defined as

$$\hat{R}(\tau) = T^{-1} \sum_{t=1}^{T-|\tau|} (y_t - \bar{y})(y_{t+|\tau|} - \bar{y}), \quad \text{where} \quad \bar{y} = T^{-1} \sum_{t=1}^{T} y_t.$$

Now, since we wish to study the behavior of the autocovariance and autocorrelation functions at all lags, as indexed by $\tau$, we consider both small-$\tau$ and large-$\tau$ asymptotics, i.e., asymptotic approximations where we have either $|\tau|/T \to 0$ or $|\tau|/T \to \kappa \in (0, 1)$ as $T \to \infty$.

Proposition A1. Suppose that $y_t$ satisfies the signal-plus-noise model in Varneskov & Perron (2017, Equations (2)-(3)), then $\hat{R}(0) = R_h(0) + \int_0^1 (J(s) - \bar{J})^2 ds + \sigma_w^2$ as $T \to \infty$ where $\bar{J} = \int_0^1 J(s) ds$. Moreover, for the cases $|\tau| > 0$, it follows that

(a) if $|\tau|/T \to 0$ as $T \to \infty$, $\hat{R}(\tau) \Rightarrow R_h(\tau) + \int_0^1 (J(s) - \bar{J})^2 ds$;

(b) if $|\tau|/T \to \kappa \in (0, 1)$ as $T \to \infty$, $\hat{R}(\tau) \Rightarrow \int_0^{1-\kappa} (J(s) - \bar{J})(J(s + \kappa) - \bar{J}) ds$.

Proof. As in Perron & Qu (2010), we can make the decomposition,

$$\hat{R}(\tau) = T^{-1} \sum_{t=1}^{T-|\tau|} (y_t - \bar{y})(y_{t+|\tau|} - \bar{y}) = T^{-1} \sum_{t=1}^{T-|\tau|} y_t y_{t+|\tau|} - \left( T^{-1} \sum_{t=1}^{T} y_t \right)^2 + o_p(1),$$

implying that we only need to provide asymptotic results for the mean and autocovariance function of the long memory component, $h_t$, for $d \in (0, 1/2)$ since the corresponding results for $d = 0$, $u_t$ and $v_t$ readily follow from the derivations of Perron & Qu (2010, pp. 288-289), who have covered the short memory case. The cross-products between the elements of $(h_t, u_t, v_t)^T$ vanish asymptotically using standard arguments due to their mutual independence. Next, under the conditions of Varneskov &
Perron (2017, Section 2), it follows by invoking Hosking (1996, Theorems 1 and 4) that

\[
T^{-1} \sum_{t=1}^{T} h_t \equiv \bar{h} = O_p \left( T^{d-1/2} \right), \quad T^{-1} \sum_{t=1}^{T-|\tau|} h_t h_{t+|\tau|} \equiv \hat{R}_h(\tau) = R_h(\tau) + O_p(f(d))
\]

where

\[
f(d) = T^{-(1-2d)} \mathbf{1}_{\{d \in (1/4, 1/2)\}} + \left( \frac{\ln T}{T} \right)^{-1/2} \mathbf{1}_{\{d = 1/4\}} + T^{-1/2} \mathbf{1}_{\{d \in (0, 1/4)\}},
\]

with \(\mathbf{1}_{\{d\}}\) being the indicator function. Hence, as \(f(d) \to 0 \forall d \in (0, 1/2)\) when \(T \to \infty\), we have both \(\bar{y} \overset{p}{\to} 0\) and \(\hat{R}_h(\tau) \overset{p}{\to} R_h(\tau)\), providing (a). The large-\(\tau\) result in (b) readily follows by combining Perron & Qu (2010, Proposition 2) with (a) since \(\hat{R}_h(\kappa T) = R_h(\kappa T) + o_p(1)\) by the small-\(\tau\) result, and for which it holds that \(R_h(\kappa T) \propto (\kappa T)^{2d-1}\), implying \(\hat{R}_h(\kappa T) \overset{p}{\to} 0\) as \(T \to \infty\).

The decomposition of the limiting autocovariance function in Proposition A1 is analogous to the result for short memory dynamics in Perron & Qu (2010) when \(\tau \neq 0\). However, it extends the latter by allowing for a hyperbolic decay in \(R_h(\tau)\). From the decomposition, we see that the contribution of the level shift process is a positive random variable, independent of \(\tau\), suggesting that \(R_h(\tau)\) will impact the total autocovariance function at first, but as \(|\tau|\) increases, the relative contribution of the level shift component will eventually dominate \(R_h(\tau) \propto |\tau|^{2d-1}\) since its limiting values does not tend to zero. However, given the possible hyperbolic decay in \(R_h(\tau)\), this may only occur at large values of \(|\tau|\). In fact, by the algebraic result,

\[
\sum_{\tau = -(T-1)}^{(T-1)} \hat{R}(\tau) = 0,
\]

see, e.g., Percival (1993), we expect that if a genuine long memory component is present in the series, the empirical autocovariance function for the log-volatility process will predominantly be negative for large \(|\tau|\) to offset the strong positive autocovariance at smaller values. Finally, Proposition A1 may readily be used to describe the autocorrelation function:

**Corollary A1.** Under the conditions of Proposition A1, denote

\[
R(\tau) = R_h(\tau) + \int_{0}^{1} (J(s) - \bar{J})^2 ds \quad \text{and} \quad R(\tau) = \int_{0}^{1-\kappa} (J(s) - \bar{J})(J(s + \kappa) - \bar{J}) ds,
\]

for \(\tau > 0\), under small-\(\tau\) and large-\(\tau\) asymptotics. Moreover, define the theoretical and empirical autocorrelation statistics \(\rho(\tau) = R(\tau)/R(0)\) and \(\hat{\rho}(\tau) = \hat{R}(\tau)/\hat{R}(0)\), respectively, as well as the a measure of the noise-to-signal ratio \(\zeta = \sigma_u^2/R(0)\), then, under both small-\(\tau\) and large-\(\tau\) asymptotics,

\[
\hat{\rho}(\tau) \Rightarrow \rho(\tau)/(1 + \zeta) \quad \text{as} \quad T \to \infty.
\]

From Corollary A1, we observe that the empirical autocorrelation functions (ACF’s) for volatility measures are downward biased, in absolute terms, with magnitudes that depend on the noise-to-signal ratio. This readily suggests that the shape of the ACF depends on whether the log-volatility proxy has
been constructed from daily or high-frequency data and, in particular, that the former likely suffers from
a more pronounced downward bias. Note that Corollary A1 extends the corresponding result in, e.g.,
Hansen & Lunde (2014, Lemma 2) by allowing for more general dependence in $\rho(\tau)$.

To gauge the qualitative features of the eight log-volatility series, their empirical ACF’s are depicted
in the top half of Figure 1 for the SPY and USD-JPY series and in Figure 2 for the six remaining series.
The top half of Figure 1 singles out the empirical ACF for the SPY and USD-JPY series since we make
a group distinction of series (SPY: BAC, MRK, S&P 500) and (USD-JPY: USD-AUD, USD-CHF), as
in the main text. The series within each group share similar characteristics and will, thus, be described
as one. The T-bond series is harder to classify as it sometimes seems to belong to the SPY group and
sometimes to the USD-JPY group. We will make the distinction clear when necessary.

The SPY series displays strong positive autocorrelation at smaller lags and negative autocorrelation at
larger lags, indicating the presence of a sizable genuine long memory component. The USD-JPY series,
on the other hand, also displays strong positive autocorrelation at smaller lags, but there is no evidence
of negative autocorrelation at larger lags, suggesting that it may be the presence of random level shifts,
in combination with a short memory component, which drives the slowly decaying autocorrelations.
Additionally, we observe that the (absolute) magnitude of the empirical ACF’s for the USD-CHF and
USD-JPY series are considerably smaller than the corresponding estimates for the remaining series, even
at smaller lags, which, as Corollary A1 suggests, may be explained by the noise-to-signal ratio for these
log-volatility proxies being larger. Finally, we stress that the presence of either genuine long memory or
random level shifts does not preclude the presence of the other, and that the ACF’s in Figures 1 and 2
only provide indicative evidence of the dynamic properties of the series.

4.2 Semi-Parametric Memory Parameter Estimates

As an initial assessment of the fractional integration order in the volatility series, we estimate $d$ using the
semi-parametric log-periodogram (GPH) estimator of Geweke & Porter-Hudak (1983), which is given
by the least-squares solution to

$$
\log I_y(\lambda_j) = c - 2d_{GPH} \log(2 \sin(\lambda_j/2)) + e_j, \quad j = 1, \ldots, m,
$$

where $\lambda_j = 2\pi j/T$ are the Fourier frequencies, and

$$
I_y(\lambda_j) = \left\langle \frac{1}{2\pi T} \sum_{t=1}^T y_t \exp(i\lambda_j t) \right\rangle^2, \quad i = \sqrt{-1},
$$

is the periodogram. We also consider the local Whittle (LW) estimator of Künsch (1987), which may
be written in profiled likelihood form as

$$
d_{LW} = \arg \min_d \log \hat{G}(d) - 2d \frac{1}{m} \sum_{j=1}^m \log \lambda_j, \quad \hat{G}(d) = \frac{1}{m} \sum_{j=1}^m \lambda_j^{2d} I_y(\lambda_j).
$$
The properties of both estimators depend on the behavior of the periodogram $I_y(\lambda_j)$ in a local neighborhood of the origin. To see this, we may readily combine results from Perron & Qu (2010) and McCloskey & Perron (2013) to obtain the decomposition

$$I_y(\lambda_j) = I_h(\lambda_j) + I_v(\lambda_j) + I_u(\lambda_j) + 2(I_{h,v}(\lambda_j) + I_{h,u}(\lambda_j) + I_{v,u}(\lambda_j)),$$

which provides information about the specific component that will dominate the periodogram at certain frequencies. In particular, random level shifts dominate the genuine long memory and noise components for frequencies satisfying $j = o(T^{(1-2d)/(2-2d)})$ and, vice versa, the genuine long memory component dominates the other two components for frequencies $j T^{(2d-1)/(2-2d)} \to \infty$. In other words, the pole near the origin is steeper for random level shifts, but taper off more quickly than the corresponding pole for the genuine long memory component, whereas the measurement noise primarily induces a bias in the periodogram at higher frequencies ordinates. Hence, the dynamic properties of the volatility series may informally be gauged by estimating the memory parameter using different values of $m$. Such estimates using either of the GPH and LW estimators are depicted in the second half of Figure 1 for the SPY and USD-JPY series and in Figure 3 for the remaining series.

For the SPY group, the memory parameter estimates converge to a number in the neighborhood of $1/2$ as $m$ increases, which, similarly to its ACF, suggests a presence of a genuine long memory component. However, the steep pole near the origin may indicate that the series also contains random level shifts. For the USD-JPY group, on the other hand, the memory parameter estimates indicate non-stationary fractional integration, as seen by the large estimate of $d$ for small values of $m$. However, as $m$ increases, the estimates gradually decrease. This pattern, as documented by Perron & Qu (2010), indicates a simultaneous presence of a random level shift component and short memory dynamics (or noise), where the effects of the latter becomes ever more important as $m$ increases, hence the decline in the estimate of $d$. Curiously, whereas the empirical ACF for the T-bond series resembles those of the SPY group, its memory parameter estimates are similar to those for the USD-JPY group.

The patterns in Figures 1-3 provide a more detailed description of the low-frequency properties of the volatility series than the preliminary log-periodogram and local Whittle estimates in the main text, which are only implemented using a bandwidth $m = \lfloor T^{1/2} \rfloor$ and which suggest that all series are fractionally integrated with $d > 1/2$, that is, in the non-stationary range.

**Remark 1.** The simultaneous presence of a steep pole for small values of the bandwidth, $m$, combined with gradually decreasing GPH and LW estimates as $m$ increases cannot be explained by the class of perturbed fractionally integrated models. The latter predicts that a stationary noise term will bias the $d$ estimates with the same sign for all values of $m$, see, for example, Deo & Hurvich (2001), Sun & Phillips (2003), and Hurvich, Moulines & Soulier (2005).
4.3 Testing for Genuine Long Memory

In the main text, we perform two tests of the null-hypothesis that the log-volatility is a genuine long memory process against the alternative of being comprised of a short memory process and random level shifts, which are based on Perron & Qu (2010) and Qu (2011). This is often referred to as testing for spurious long memory. The two tests are performed in the frequency domain, and they are essentially designed to utilize the insights from the decomposition in (7) about the behavior of the periodogram for a genuine long memory process and random level shifts, respectively, at different frequency ordinates.

The first test, proposed by Perron & Qu (2010), is computed as

\[ S_d(a, b) = \sqrt{\frac{24}{\pi^2}} \left( d_{GPH}(a) - d_{GPH}(b) \right) \xrightarrow{D} N(0, 1) \]

for \(0 < a < b < 1\) where \(d_{GPH}(a)\) is the GPH estimate using \(m = \lfloor T^a \rfloor\) frequency ordinates. In particular, we follow Perron & Qu (2007) and implement the test with \(b = 4/5\) and \(a = (1/3, 1/2)\).

The second test, proposed by Qu (2011), is a score-based statistic from the profiled Whittle likelihood function, defined as

\[ W_\epsilon = \sup_{r \in [\epsilon, 1]} \left( \sum_{j=1}^m \bar{\omega}_j^2 \right)^{-1/2} \left| \sum_{j=1}^{m\gamma} \bar{\omega}_j \left( \frac{I_\gamma(\lambda_j)}{G(d_{LW})\lambda_j^{-2d_{LW}}} - 1 \right) \right| \]

where \(\bar{\omega}_j = \ln \lambda_j - m^{-1} \sum_{j=1}^m \ln \lambda_j\) and \(\epsilon\) is a trimming parameter. We implement the test using trimming \(\epsilon = (0.02, 0.05)\), for which the critical values are provided in Qu (2011, Table 1), and a LW estimate using \(m = \lfloor T^{0.7} \rfloor\) frequency ordinates. The \(W_\epsilon\) test has important advantages over the \(S_d(a, b)\) testing procedure. First, it allows the process under the alternative to be comprised of a genuine long memory component and random level shifts, which corresponds well with the proposed reduced form model in Varneskov & Perron (2017, Equation (4)) where both components are allowed. Second, it is more robust against measurement errors and short memory dynamics since it is implemented with a (consistent) pre-whitening procedure, see Qu (2011, Section 5). Finally, Qu (2011) shows that the test has good size and power properties for sample sizes \(T \geq 2000\).

5 Maximum Likelihood Estimation

The basic principle behind the estimation procedure is to augment the probability of states (or level regimes) by the realizations of a mixture of normally distributed processes at time \(t\) and apply the Kalman filter to construct the likelihood function conditional on the realization of states. Before proceeding, denote the observations up to and including time \(t\) by \(Y_t = (\Delta y_2, \Delta y_3, \ldots, \Delta y_t)'\) and let us collect the model parameters in a vector \(\Sigma = (\sigma_y, \gamma, \sigma_\varepsilon, d, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'\). Then, we may express
the conditional log-likelihood function as

\[
\ln(L) = \sum_{i=1}^{T} \ln f(\Delta y_t|Y_{t-1}; \Sigma)
\]

\[
f(\Delta y_t|Y_{t-1}; \Sigma) = \sum_{i=0}^{1} \sum_{j=0}^{1} f(\Delta y_t|s_{t-1} = i, s_t = j, Y_{t-1}; \Sigma) \Pr(s_{t-1} = i, s_t = j|Y_{t-1}; \Sigma)
\]

where \( s_t = \pi_{T,t} \) is used as an indicator for the particular state at time \( t \), which is independent of past realizations. As such, the likelihood function resembles its counterpart for Markov regime switching models, see, e.g., Hamilton (1994), but, as will become apparent below, it has two added complexities relative to the estimation of such models. First, the mean and the variance of the conditional density are nonlinear functions of past realizations and the fundamental parameters. Hence, we cannot separate all elements of \( \Sigma \) using first-order conditions and apply a standard EM algorithm. Second, the conditional probability of being in a given regime is not separable from the conditional density.

To develop a feasible estimation algorithm, we first summarize some well-known rules and expressions for conditional probabilities that apply to the present setting:

\[
\Pr(s_{t-1} = i, s_t = j|Y_{t-1}; \Sigma) = \Pr(s_t = j) \Pr(s_{t-1} = i|Y_{t-1}; \Sigma)
\]

\[
= \Pr(s_t = j) \sum_{k=0}^{1} \Pr(s_{t-2} = k, s_{t-1} = i|Y_{t-1}; \Sigma),
\]

\[
\Pr(s_{t-2} = k, s_{t-1} = i|Y_{t-1}; \Sigma) = \frac{f(\Delta y_{t-1}|s_{t-2} = k, s_{t-1} = i, Y_{t-2}; \Sigma)}{f(\Delta y_{t-1}|Y_{t-2}; \Sigma)} \Pr(s_{t-2} = k, s_{t-1} = i|Y_{t-2}; \Sigma),
\]

where the second equality follows from \( s_t \) being independent of past realizations and

\[
\Pr(s_t = j|Y_{t}; \Sigma) = \sum_{i=0}^{1} \Pr(s_{t-1} = i, s_t = j|Y_{t}; \Sigma).
\]

The third equation follows by Bayes’ rule. Next, define the prediction error as

\[
\nu_t^{ij} = \Delta y_t - \mathbb{E}[\Delta y_t|s_{t-1} = i, Y_{t-1}; \Sigma] = \Delta y_t - FH_{t|t-1}^{i}, \quad \nu_t^{ij} \sim \mathcal{N}(0, f_t^{ij})
\]

whose state-dependent variance is given by \( f_t^{ij} = FP_{t|t-1}^{i}F' + \sigma_{ij}^2 \) where \( P_{t|t-1}^{i} \) is the conditional covariance matrix of the state vector in state \( i \). The generic superscript \((ij)\) refers to a matrix-valued variable conditional on the process being in state \( i \) at time \( t - 1 \), and state \( j \) at time \( t \). Note that the conditional expectation of \( \Delta y_t \) does not depend on the value of \( j \) since we are conditioning on the available information at time \( t - 1 \). In this setting, the best forecast of the state vector and its associated covariance matrix may be written as

\[
H_{t|t-1}^{i} = GH_{t-1|t-1}^{i}, \quad P_{t|t-1}^{i} = GP_{t-1|t-1}^{i} G' + Q
\]
where $H_{t-1|t-1}^i$ and $P_{t-1|t-1}^i$ are computed using standard Kalman updating equations. Specifically, the updating formulas for $s_{t-1} = i$ and $s_t = j$ are given by
\begin{align*}
H_{t|t}^{ij} &= H_{t-1|t-1}^i + P_{t-1|t-1}^i F'(FP_{t-1|t-1}^i F' + \sigma'^2_{ij})^{-1} \nu_{t}^{ij}, \\
P_{t|t}^{ij} &= P_{t|t}^i - P_{t|t-1}^i (FP_{t|t-1}^i F' + \sigma'^2_{ij})^{-1} FF_{t-1}^i
\end{align*}

where a problem arises since the two possible states cause the number of estimates for the state vector and its conditional covariance matrix to grow over time with a factor $t^2$. A solution to this, suggested in Harrison & Stevens (1976), is to re-collapse $H_{t|t}^{ij}$ and $P_{t|t}^{ij}$ to make them unaffected by the history of states before time $t - 1$ as follows
\begin{align*}
H_{t|t}^{ij} = \frac{\sum_{i=0}^{1} \Pr(s_{t-1} = i, s_t = j|Y_t; \Sigma) H_{t|t}^{ij}}{\Pr(s_t = j|Y_t; \Sigma)}, \\
P_{t|t}^{ij} = \frac{\sum_{i=0}^{1} \Pr(s_{t-1} = i, s_t = j|Y_t; \Sigma) \left[P_{t|t}^{ij} + \left(H_{t|t}^{ij} - H_{t|t}^{ij}\right) \left(H_{t|t}^{ij} - H_{t|t}^{ij}\right)'\right]}{\Pr(s_t = j|Y_t; \Sigma)},
\end{align*}

utilizing the rules for conditional probabilities provided above. Finally, combine all previous equations to express the conditional density as
\begin{equation}
\begin{aligned}
f(\Delta y_t|s_{t-1} = i, s_t = j, Y_{t-1}; \Sigma) &= \frac{1}{\sqrt{2\pi}} \left(f_{t}^{ij}\right)^{-\frac{1}{2}} \exp \left\{-\frac{\nu_t^{ij}\left(f_t^{ij}\right)^{-1/2}}{2}\right\}.
\end{aligned}
\end{equation}

As mentioned above, one complexity in the estimation arises since $\mathbb{E}[\Delta y_t|s_{t-1} = i, Y_{t-1}; \Sigma]$ and the prediction error $f_{t}^{ij}$ are non-linear functions of $\Sigma$, $Y_{t-1}$ and the conditional density (8). The second complexity is caused by (8) not being separable from $\Pr(s_{t-1} = i, s_t = j|Y_{t-1}; \Sigma)$ since the latter enters into the construction of the former through $H_{t|t}^{ij}$ and $P_{t|t}^{ij}$.

From this exhibition, we also see that if either $\gamma = 0$ or $\sigma_\eta = 0$, the other parameter is not identified and the above estimation algorithm collapses to the genuine long memory state space framework of Chan & Palma (1998). In this case, from their Theorems 3.1 and 3.2, the state space estimator of the ARFIMA parameters $\Pi = \Sigma \setminus \{\gamma, \sigma_\eta\}$, denoted $\hat{\Pi}$, is consistent as $T \to \infty$ as $M = T^\beta$ with $\beta > 0$; and when $\beta \geq 1/2$, $\sqrt{T} (\hat{\Pi} - \Pi) \overset{D}{\to} N(0, \Lambda^{-1}(\Pi))$ where $\Lambda(\Pi)$ is the usual information matrix. In other words, the ARFIMA parameters are estimated with the usual maximum likelihood properties, and they unaffected by the possibility of non-identification of the random level shift parameters.

## 6 The HAR and GARCH Models

The HAR model has been shown by, e.g., Andersen, Bollerslev & Diebold (2007), Corsi (2009), and Chiriac & Voev (2011) to provide accurate forecasts of realized volatility measures. It is a regression-based approximate long memory model, which captures the hyperbolically decaying autocorrelations of
persistent time series by weighting lagged AR terms in a parsimonious way. In particular, it may be written as

\[ y_{t+1}^{(d)} = \alpha + \beta_1 y_t^{(d)} + \beta_2 y_t^{(w)} + \beta_3 y_t^{(bw)} + \beta_4 y_t^{(m)} + \epsilon_{t+1}^{(d)} \]

where \( d, w, bw, \) and \( m \) denote a daily, weekly (5 days), biweekly (10 days), and monthly (21 days) sampling frequency, respectively, \( \alpha \) is a constant and \( \epsilon_t^{(d)} \sim i.i.d.N(0,\sigma^2) \). The regressors on the right-hand-side are averages of past values of \( y_t \) scaled to match the left-hand-side variable, for example, for the weekly term \( y_t^{(w)} = \frac{1}{5} \sum_{j=0}^{4} y_{t-j} \). The model is estimated using maximum likelihood, which is equivalent to OLS in this setting. Direct out-of-sample forecasts from the HAR model are easily obtained due to the hierarchical structure of the model. A one-step-ahead forecast may be computed from the model above, while multi-step-ahead forecasts are constructed by specifying the hierarchy to match the forecast horizon, see, e.g., Chiriac & Voev (2011) for details.

The GARCH model is implemented using the discrete or continuously compounded return model in (1) or (4), respectively, depending on whether we have daily or high-frequency data available to construct the volatility proxy. Both return models, however, imply that we may adopt a standard Gaussian likelihood function for estimation of the model parameters. For the latent discrete time log-volatility process in the signal-plus-noise model in Varneskov & Perron (2017), \( x_t \), we consider a GARCH(1,1) specification; see, e.g., Bollerslev (1986). As the GARCH model provides an explicit filter of the measurement errors in the volatility proxies, we may write \( x_t \) in terms of the observable proxy, \( y_t \), as follows

\[ x_t = \alpha + \beta_1 x_{t-1} + \beta_2 y_{t-1}. \]

From the extrapolated model parameters, we construct \( \tau \)-step-ahead forecasts of \( x_t \) according to

\[ \hat{x}_{t+1|t} = \alpha + \beta_1 x_t + \beta_2 y_t, \quad \hat{x}_{t+\tau|t} = \varphi + (\beta_1 + \beta_2)^{-1}(\hat{x}_{t+1|t} - \varphi), \quad \tau > 1, \]

where \( \varphi = \alpha/(1-\beta_1-\beta_2) \), see, e.g., Andersen, Bollerslev, Christoffersen & Diebold (2006). Interestingly, however not surprisingly, the estimated loading on the innovations, or news, to the (latent) log-volatility process, \( \beta_2 \), differs quite suggestively across the two groups; the high-frequency measures in the SPY group receive a loading around 1/2, whereas it is much smaller for the USD-JPY group and the T-bond series. This suggests that high-frequency measures of volatility are more informative about future volatility than measures constructed from daily data.

7 The Model Confidence Set

Consider a set \( M^0 \) that contains a finite number of objects indexed by \( i = 1, \ldots, m_0 \). These objects are evaluated in terms of a loss function, \( L \), over the sample \( t = 1, \ldots, T \), where the loss associated with period \( t \) is denoted by \( L_{i,t} \). Define the relative performance \( d_{ij,t}, \forall i, j \in M^0 \). Then, the set of superior

---

5This could be, e.g., mean squared forecast errors or mean absolute forecast errors.
The objective is to determine \( \mathcal{M}^* \), and this is done through a sequence of significance tests where the significantly inferior objects of \( \mathcal{M}_0 \) are eliminated. The null hypothesis may be stated as
\[
H_{0,\mathcal{M}} : \mathbb{E}[d_{ij,t}] = 0 \quad \forall i, j \in \mathcal{M} \subset \mathcal{M}_0,
\]
which is tested against the alternative \( \mathbb{E}[d_{ij,t}] \neq 0 \) for some \( i, j \in \mathcal{M} \). The test is based on an algorithm that consists of an equivalence test, \( \delta_{\mathcal{M}} \), and an elimination rule, \( e_{\mathcal{M}} \). The equivalence test is used to assess \( H_{0,\mathcal{M}} \), and it takes values \( \delta_{\mathcal{M}} = \{0, 1\} \) corresponding to accepting or rejecting \( H_{0,\mathcal{M}} \), respectively.

The elimination rule, \( e_{\mathcal{M}} \), determines the object of \( \mathcal{M} \) that is to be removed in the event that \( H_{0,\mathcal{M}} \) is rejected. The MCS algorithm may, thus, be described by three steps:

**Step 1** Initially set \( \mathcal{M} = \mathcal{M}_0 \).

**Step 2** Test \( H_{0,\mathcal{M}} \) using \( \delta_{\mathcal{M}} \) at a given significance level \( \alpha \).

**Step 3** If \( \delta_{\mathcal{M}} = 0 \), define \( \hat{\mathcal{M}}^{*}_{1-\alpha} = \mathcal{M} \); otherwise use \( e_{\mathcal{M}} \) to eliminate an object from \( \mathcal{M} \) and repeat the procedure from Step 1.

The set \( \hat{\mathcal{M}}^{*}_{1-\alpha} \) consists of the surviving objects, and this is referred to as the model confidence set.

### 7.1 MCS p-values

To facilitate the interpretation of the p-values, consider the sequence of random sets \( \mathcal{M}_0 = \mathcal{M}_1 \supset \mathcal{M}_2 \supset \cdots \supset \mathcal{M}_m \) where \( \mathcal{M}_i = \{e_{\mathcal{M}_i}, \ldots, e_{\mathcal{M}_m}\} \), so that \( e_{\mathcal{M}_1} \) is the first element to be eliminated in the event that \( H_{0,\mathcal{M}_1} \) is rejected, \( e_{\mathcal{M}_2} \) is the second element, and so on. Then, let \( p_{H_{0,\mathcal{M}_i}} \) denote the p-value associated with the null hypothesis \( H_{0,\mathcal{M}_i} \), with the convention \( p_{H_{0,\mathcal{M}_0}} \equiv 1 \). The MCS p-value for model \( e_{\mathcal{M}_j} \in \mathcal{M}_0 \) is defined in this setup by \( \hat{p}_{e_{\mathcal{M}_j}} \equiv \max_{i \leq j} p_{H_{0,\mathcal{M}_i}} \).

The interpretation of the MCS p-values are analogous to that of standard p-values. The MCS set may, thus, be interpreted as containing the best random subset of models, \( \mathcal{M}^* \), with a certain probability.

### 7.2 Equivalence Test and Elimination Rule

Several equivalence tests and elimination rules have been suggested, see Hansen et al. (2011). For the empirical implementation, we have selected the range statistic, \( T_{R,\mathcal{M}} \equiv \max_{i,j \in \mathcal{M}} |t_{ij}| \), where \( t_{ij} \) is a t-statistic constructed as
\[
t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{\text{Var}(\bar{d}_{ij})}} \quad \text{for} \quad i, j \in \mathcal{M}
\]
where \( \bar{d}_{ij} = \frac{1}{n} \sum_{t=1}^{T} d_{ij,t} \), i.e., the average relative loss between the \( i \)th and \( j \)th models and \( \text{Var}(\bar{d}_{ij}) \) is its variance. The elimination rule is, then, given by \( e_{R,\mathcal{M}} = \arg \max_{i \in \mathcal{M}} \sup_{j \in \mathcal{M}} t_{ij} \).
8 Proof of Proposition 1

Following the Kalman filter iterations and the Harrison & Stevens (1976) re-collapsing procedure, the best one-step ahead forecast of $\Delta y_{t+1}$ will only depend on four transition paths, indicated by the superscript $(ij)$, as follows

$$E_t[\Delta y_{t+1}] = F \sum_{i=0}^{1} \sum_{j=0}^{1} \Pr(\pi_{T,t} = i, \pi_{T,t+1} = j | Y_t; \Sigma) \mathbb{E}_t \left[ H_{t+1|t+1}^{ij} \right],$$

Now, by applying the definition of the prediction error in the recursive estimation algorithm described above in Section 5, $\nu_{t+1}^{ij} = \Delta y_{t+1} - E[\Delta y_{t+1} | \pi_{T,t} = i, Y_t; \Sigma]$ and $E_t[\nu_{t+1}^{ij}] = 0$, in conjunction with the Kalman updating equation $E_t[H_{t+1|t+1}^{ij}] = H_{t+1|t}^{ij} = GH_{t|t}^{ij}$, we have

$$E_t[\Delta y_{t+1}] = FG \sum_{i=0}^{1} \sum_{j=0}^{1} \Pr(\pi_{T,t} = i, \pi_{T,t+1} = j | Y_t; \Sigma) H_{t|t}^{ij},$$

where the re-collapsing procedure of Harrison & Stevens (1976) is not adopted in the last step, since our forecasts only depend on four transition paths. We may, then, obtain a $\tau$-step-ahead prediction using the recursive algorithm for best linear mean-square predictors (see, e.g., Brockwell & Davis (1991) for details) and the fact that the probability of a future random level shift is assumed to be independent of past information, $Y_t$, that is,

$$\Pr(\pi_{T,t+\tau} = j | \pi_{T,t} = i, Y_t; \Sigma) = \Pr(\pi_{T,t+\tau} = j) = \Pr(\pi_{T,t+1} = j), \quad \tau \geq 1.$$ 

As a result, since $F^\tau = F$, we may express the $\tau$-step-ahead forecast as

$$E_t[y_{t+\tau}] = y_t + FG^\tau \sum_{i=0}^{1} \sum_{j=0}^{1} \Pr(\pi_{T,t+\tau} = j) \Pr(\pi_{T,t} = i | Y_t; \Sigma) H_{t|t}^{ij},$$

where all components on the right-hand-side may be obtained using the information available in the updating equations in Section 5, concluding the proof.

9 Proof of Proposition 2

Since, conditional on time $t$ information, $\nu_{t+1}^{ij} \sim N(0, f_{t+1}^{ij})$ is Gaussian and independent over time, we know that

$$E_t[\exp(y_{t+\tau})] = \exp \left( E_t[y_{t+\tau}] + \nabla_t[y_{t+\tau}] / 2 \right).$$

Moreover, as we have $E_t[y_{t+\tau}] = \hat{y}_{t+\tau|t}$, with $\hat{y}_{t+\tau|t}$ provided by Proposition 1, we are left with proving the explicit form for $\hat{\zeta}_{t+\tau|t}$, as stated in the proposition. Hence, by the telescoping sum property for log
innovations, write

\[ \mathbb{V}_t[\Delta y_{t+s}] = \mathbb{E}_t[(\Delta y_{t+s} - \mathbb{E}[\Delta y_{t+s}])^2] = \sum_{s=1}^{\tau} \mathbb{E}_t[(\Delta y_{t+s})^2] = \sum_{s=1}^{\tau} \mathbb{V}_t[\Delta y_{t+s}], \]

using conditional independence of \( \nu_{t+k}^{ij} \) and \( \nu_{t+l}^{ij} \) for \( k \neq l \) and \( k, l \geq 0 \) for the second equality. By applying the updating equations in Section 5 without using the Harrison & Stevens (1976) re-collapsing procedure, as for Proposition 1, we may write

\[ \mathbb{V}_t[\Delta y_{t+s}] = \sum_{i=0}^{1} \sum_{j=0}^{1} \text{Pr}(\pi_{T,t+s} = j) \text{Pr}(\pi_{T,t} = i|Y_t; \Sigma) \mathbb{V}_t[\nu_{t+s}^{ij}] \]

\[ = \sum_{i=0}^{1} \sum_{j=0}^{1} \text{Pr}(\pi_{T,t+s} = j) \text{Pr}(\pi_{T,t} = i|Y_t; \Sigma) \left( FG^{s-1} \left( GP^{ij}_{t|t} \Sigma' + Q \right) \left( FG^{s-1} \right)' + \sigma_{ij}^2 \right), \]

providing the final result. \( \square \)
Autocorrelation Function: SPY HF

Autocorrelation Function: USD-JPY

Semiparametric Memory Estimate: SPY HF

Semiparametric Memory Estimate: USD-JPY

Figure 1: The upper half shows autocorrelation functions for the first $T/2$ lags. The lower half shows log-periodogram and local Whittle estimates of $d$ as a function of the number of frequency ordinates used.
References


Figure 2: Autocorrelation functions for the first $T/2$ lags.
Figure 3: Log-periodogram and local Whittle estimates of $d$ as a function of the number of frequency ordinates used.


