Working Paper (6 January 2022)

Volatility Puzzle

Jun Yu
Singapore Management University

Shuping Shi
Macquarie University

SMU
SINGAPORE MANAGEMENT UNIVERSITY

SIM KEE BOON INSTITUTE FOR FINANCIAL ECONOMICS

LEE KONG CHIAN SCHOOL OF BUSINESS

Volatility Puzzle[*]

Shuping Shi
Macquarie University

Jun Yu
Singapore Management University

January 6, 2022

Abstract

The log realized volatility (RV) is often modeled as an autoregressive fractionally integrated moving average model (ARFIMA$(1, d, 0)$). Two conflicting empirical results have been found in the literature. One stream shows that log RV has a long memory (i.e., the fractional parameter $d > 0$). The other stream suggests that the autoregressive coefficient α is near unity with anti-persistent errors (i.e., $d < 0$). This paper explains how these conflicting empirical findings can co-exist in the context of ARFIMA$(1, d, 0)$ model by examining the finite sample properties of popular estimation methods, including semi-parametric methods and parametric maximum likelihood methods. The finite sample problems suggest that it is difficult to distinguish ARFIMA$(1, d, 0)$ with α close to zero and d close to 0.5 from ARFIMA$(1, d, 0)$ with α close to unity and d close to -0.5 as both models are related to a unit root model with anti-persistent errors. For the ten financial assets considered, despite no definitive conclusions can be drawn regarding the data generating process, we find that the frequency domain maximum likelihood (or Whittle) method can generate the most accurate out-of-sample forecasts.

JEL classification: C15, C22, C32
Keywords: Long memory; fractional integration; roughness; short-run dynamics; realized volatility

[*] We are grateful to Dacheng Xiu for providing us his realized volatility data, Mikkel Bennedsen, Kim Christensen, Jia Li, Morten Nielsen, Peter Phillips, an associate editor, and two referees for constructive comments. Shi acknowledges research support from the Australian Research Council under project No. DE190100840. Yu would like to acknowledge that this research/project is supported by the Ministry of Education, Singapore, under its Academic Research Fund (AcRF) Tier 2 (Award Number MOE-T2EP402A20-0002). He also acknowledges the financial support from the Lee Foundation. Shuping Shi, Department of Economics, Macquarie University; E-mail: shuping.shi@mq.edu.au. Jun Yu, School of Economics and Lee Kong Chian School of Business, Singapore Management University, 90 Stamford Rd, Singapore 178903. Email: yujun@smu.edu.sg.

1 Introduction

The availability of intraday prices of financial assets fosters the development of high-frequency financial econometrics, making an accurate measurement of daily 'realized' volatility possible. The estimated daily realized volatility (RV) has been shown helpful for various purposes, including forecasting macroeconomic fundamentals (Andersen, Bollerslev, Diebold, and Wu, 2005), making investment decisions (Fleming, Kirby, and Ostdiek, 2003), pricing options (Christoffersen, Feunou, Jacobs, and Meddahi, 2014), managing financial risk (Christoffersen and Diebold, 2000), and estimating model parameters (Phillips and Yu, 2009; Tao, Phillips, and Yu, 2019).

A class of autoregressive fractionally integrated moving average (ARFIMA) models, particularly ARFIMA(p, d, q) with $p = 1$ and $q = 0$, has gained much prominence in modeling daily log RV. For notational convenience, in the rest of this paper, we denote RFIMA$(1, d, 0)$ with the autoregressive coefficient α by AR1FI(α, d). When $d > 0$, the autocorrelation function (ACF) of AR1FI(α, d) decays hyperbolically and is not absolutely summable. This feature matches well with the empirical ACF observed in data. The value of the fractional parameter d has important implications for both the theoretical and empirical analysis of RV. As such, the main focus of the literature has been on the estimation of d.

Several estimation techniques for d have been proposed, including the local Whittle estimation (LWE) method (Künsch, 1987; Robinson, 1995a) and the log periodogram estimation (LPE) method (Geweke and Porter-Hudak, 1983; Robinson, 1995b). These two methods rely on the asymptotic behavior of the spectral density at frequencies near zero (ignoring short-run dynamics) and hence, are often referred to as semi-parametric methods. When the two semi-parametric methods are applied to log RV, it is often found that the point estimate of d is around 0.5. See, for example, Andersen and Bollerslev (1997), Andersen, Bollerslev, Diebold, and Labys (2001), Andersen, Bollerslev, Diebold, and Ebens (2001), Andersen, Bollerslev, Diebold, and Labys (2003), Baillie, Calonaci, Cho, and Rho (2019). Such an estimate implies that the log RV has a long memory.[1] The presence of long memory in RV has been widely regarded as a stylized fact. Furthermore, one could estimate the short-run parameter from the pre-filtered data (based on the estimated d). The estimated short-run parameter typically suggests weak short-run behavior or strong mean reversion. When an AR(1) model is fitted to the filtered data, the estimated autoregressive parameter α is often close to zero. For convenience, we label the AR1FI(α, d) process with α close to zero and d near 0.5 by Model 1.

One advantage of the semi-parametric methods is their asymptotic robustness to short-run dynamics, as short-run behavior does not change the asymptotic spectral density at near-zero frequencies. This insensitive relationship, however, does not necessarily hold in finite samples. In particular, AR1FI(α, d) with α close to unity is similar to AR1FI$(0, d + 1)$, and the spectral density that ignores the near-unity behavior is expected to approximate the actual spectral density poorly, even with a large sample size, at frequencies near zero.[2] This concern might have important empirical implications.

It is known that the ARFIMA$(0, d, 0)$ model with $d \in (0, 1/2)$ is asymptotically equivalent to the frac-

[1] Long memory is typically defined within the class of stationary models and refers to the case of $d \in (0, 1/2)$. Our definition of long memory here is broader. It refers to the case od $d > 0$ as in Phillips and Shimotsu (2004).

[2] In fact, AR1FI$(0, d)$ is observationally equivalent to AR1FI$(1, d - 1)$.

tional Gaussian noise (fGn) with the Hurst parameter H with $H = d + 0.5$. The fGn is the increment of the fractional Brownian motion (fBm), denoted by $B^H(t)$, whose sample path is (locally) Hölder continuous up to order H. Based on fBm, Wang, Xiao, and Yu (2021) consider a fractional Ornstein-Uhlenbeck (fOU) process for log RV. Under the in-fill asymptotic scheme, the discrete-time representation of the fOU process is a local-to-unity (Phillips, 1987) process with fGn, which is asymptotically equivalent to AR1FI(α, d) with α close to unity and $d = H - 0.5$.[3] A change-of-frequency method is proposed to estimate H in fOU in Wang, Xiao, and Yu (2021). The estimated H from several log RV series suggests $H < 0.5$ (i.e. $d < 0$). Similar empirical estimates of H are found in Bolko, Christensen, Pakkanen, and Veliyev (2021) when the fOU model is used to capture the movement of log spot volatility and the generalized moment of method is used to estimate H. The (pre-imposed) local-to-unity dynamic of the fOU model generates strong persistency that is attenuated by the anti-persistent errors. We label the AR1FI(α, d) process with α close to unity and $d < 0$ by Model 2.

Clearly, the empirical evidence for log RV by semi-parametric methods is at odds with that obtained from the fOU model. While the semi-parametric methods suggest a process with a weak short-run dynamic and long memory errors (i.e. Model 1) for log RV, the empirical evidence obtained from the fOU model reveals near-unity behavior and anti-persistent errors (i.e., Model 2). We are concerned with this volatility puzzle in the present paper.

The first goal of this paper is to understand how these conflicting empirical findings co-exist in the literature. To achieve this goal, we examine the finite sample properties of several popular estimation methods for AR1FI(α, d) under a wide range of parameter settings. The methods include two semi-parametric methods and two parametric maximum likelihood (ML) methods. The semi-parametric methods are LWE applied to log RV and LWE applied to the first difference of log RV, later of which is referred to as LWE(diff) hereafter. The two ML approaches are the modified profile time-domain likelihood (MPL) method and the frequency domain maximum likelihood (Whittle) method. Both classes of estimation approaches have some finite sample problems under one or both of Model 1 and 2.

For the semi-parametric methods, it is found that when the true data generating process (DGP) is Model 2, LWE points to Model 1. When the true DGP is Model 1, LWE(diff) points to AR1FI($1, d$) with d being negative. These findings hold true even when the sample size is very large in an empirically realistic situation. Moreover, the LWE (LWE(diff)) estimator is substantially biased when the autoregressive coefficient deviates far from zero (unity).[4] In contrast, the two parametric ML methods generally perform well. However, it is possible for both MPL and Whittle to mix up Model 1 and 2 in finite samples. Specifically, when the DGP is Model 1, with a small and non-negligible probability, both methods lead to Model 2. On the other hand, when the true DGP is Model 2, the parametric ML methods could point to Model 1. These problems arise because there are two modes in their likelihood

[3] Similar models have been considered in other papers. For example, Magdalinos (2012) proposes a mildly explosive autoregressive process with a long memory errors (i.e., $d \in (0, 0.5)$). Yu (2021) considers a latent local-to-unity model with fractionally integrated errors.

[4] Unreported simulations show that results remain the same when using other popular semi-parametric methods (such as LPE and the exact local Whittle method of Shimotsu and Phillips (2006)) or when tapering (Dahlhaus, 1988; Velasco, 1999) is applied. The tapering technique employed is described in the appendix.

functions, and the mode around the true parameter values may be lower than the other mode in finite samples. As a result, the finite sample distribution can be bi-modal. The simulation findings calls for cautious interpretation of empirical estimation results when using those techniques.

We consider log RV time series of ten financial assets spanning over a decade from 2010 to 2021. The estimates from LWE lead to Model 1 ($\alpha \in [-0.162, 0.004]$ and $d \in [0.54, 0.70]$), while the two ML methods point to Model 2 ($\alpha \in [0.995, 0.999]$ and $d \in [-0.47, -0.38]$). Results of LWE(diff) are close to those from the ML methods but with $\alpha = 1$ by assumption. Both LWE and LWE(diff) suggest nonstationarity, whereas, by assumption, the estimated processes from two ML methods are stationary. Since the ML methods are relatively reliable, Model 2 is more likely to be the true model than Model 1. Nevertheless, there is still a small chance that Model 1 is the true DGP. Despite the inconclusive estimation results, we show that the Whittle method can provide the best out-of-sample forecast out of the four estimation techniques (especially at long forecasting horizons), followed by MPL.

Our paper contributes to the literature in two aspects. First, our simulation findings explain how Model 1 and Model 2 co-exist in the RV literature. While the simulation studies in the existing literature[5] have found a substantial upward bias in d with the semi-parametric methods when α takes a large positive value and examined the performance of the ML methods under various parameter settings, the simulation designs adopted in these studies prevent them from finding the difficulty of the two classes of estimation methods in distinguishing Model 1 from Model 2. In particular, the selected values for α are too far away from unity so that the bias generated by semi-parametric methods is not substantial enough and that ML methods can well distinguish Model 1 from Model 2. We find that under Model 1 and 2, the finite sample distributions of the two ML methods are bi-modal. One mode is in the parameter ranges of Model 1 and the other one corresponds to Model 2. Consequently, the traditional summary statistics of the estimates such as mean (or the bias) and standard errors (or root mean squared errors) are poor choices. Alternative measures are proposed. Second, although we cannot draw definitive conclusions regarding the DGP with the estimation methods in empirical applications due to their finite sample issues, we find that the Whittle method provides the best out-of-sample forecasts out of the four.

The paper is organized as follows. Section 2 introduces the RV estimator. Section 3 presents the model specification and reviews some statistical properties of the model. Section 4 introduces the four popular estimation methods. Section 5 presents the simulation designs and reports the finite sample properties of the estimation approaches. Section 6 reports empirical estimation and forecasting results. Section 7 concludes. The Appendix reviews a technique, known as tapering, for LWE and Whittle. We also examine the robustness of our empirical results using alternative volatility measures in the appendix.

[5] See Smith, Taylor, and Yadav (1997), Nielsen and Frederiksen (2005), and Nadarajah, Martin, and Poskitt (2021).

2 Realized Volatility

Assume data are observed at a regular frequency. Let $t = 1, \cdots, T$ and $n = T/\delta$ be the total number of intra-day observations available within sample period, where δ is the distance between two consecutive observations. Let $X_{t,i}$ be the observed i^{th} log prices at period t. The traditional realized volatility is constructed as

$$RV_t = \sum_{i=2}^{1/\delta} \left(\Delta X_{t,i}\right)^2, \text{ with } \Delta X_{t,i} = X_{t,i} - X_{t,i-1}. \tag{1}$$

Under a standard Itó-semimartingale process, the realized volatility is shown to be a consistent estimator of the quadratic variation of the process.

One of the most recent contributions in the volatility estimation literature is made by Da and Xiu (2021), who develop a quasi-ML (QML) approach providing uniform valid inference on volatility under an extremely general model setting with both MA(∞) market microstructure noises and jumps.[6] The model specification considered by Da and Xiu (2021) is as follows. The observed log asset prices consist of two components:

$$X_t^o = X_t + U_t,$$

where X_t is the underlying log efficient price and U_t is the noise component. The noise process U_t is assumed to have flexible serial correlations, modeled as an MA(∞) process. The underlying price is assumed to be an Itó-semimartingale process defined on some filtered probability space $\left(\Omega, \mathcal{F}, \left(\mathcal{F}_t\right), \mathbb{P}\right)$ and satisfies

$$X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + \left(\delta 1_{\{|\delta| \leq 1\}} * (\eta - \upsilon)\right)_t + \left(\delta 1_{\{|\delta|>1\}}\right) \mu_t, \tag{2}$$

where μ_t and σ_t are adapted and locally bounded, W is a standard Brownian motion, η is a Poisson random measure on $\mathbb{R}_+ \times E$ with a non-random intensity measure $\upsilon\left(dt, ds\right) = dt \otimes \lambda\left(ds\right)$, and λ is a σ-finite measure on (E, ξ) which is a Polish space. The last two components of (2) capture the dynamics of jumps. See, for example, Jacod, Li, and Zheng (2017) or Da and Xiu (2021) for more details of the assumptions.

The likelihood function of QML is taken from a much simplified process, assuming the efficient price follows a Brownian motion with constant volatility and a Gaussian MA(q) noise component. The QML estimator of the volatility, denoted by $\hat{\sigma}^2\left(\hat{q}\right)$ with $\hat{q}$ obtained from the Akaike information criterion, is shown to converges to the following quadratic variation,

$$\delta \left[\int_{t-1}^t \sigma_s^2 ds + \sum_{i=2}^{1/\delta} \left(\Delta X_{t,i}\right)^2\right],$$

which comprises both continuous (integrated variance) and discontinuous (jump) components.

[6] Other noise-robust volatility estimators include the traditional RV obtained from returns sampled at the 5-minute frequency, the pre-averaging method of Jacod, Li, Mykland, Podolskij, and Vetter (2009), and the flat-top realized kernel estimator of Varneskov (2017).

This paper investigates the dynamics of the QML volatility estimates of various financial assets. The QML realized volatility data are conveniently provided by the Risk Lab[7] and computed using transaction prices sampled at the highest available frequency. For convenience, we refer to the QML volatility estimator as log RV (QML) subsequently and, with a slight abuse of notation, we denote the log RV (QML) volatility estimator $\log \hat{\sigma}^2 (\hat{q})$ by y_t.

3 Model Specification

Consider the AR1FI(α, d) model

$$(1 - \alpha L) (y_t - \mu) = \sigma_u u_t, \tag{3}$$

where L is the lag operator, and u_t is the error term. The error term is a fractionally integrated process (Granger and Joyeux, 1980) such that

$$u_t = (1 - L)^{-d} \varepsilon_t \text{ with } \varepsilon_t \sim_{iid} N(0, 1), \tag{4}$$

where d is the memory parameter. The AR1FI(α, d) model is one of the most popular specifications for modeling log RV in the literature. See, for example, Andersen et al. (2003) and Wang et al. (2021).[8] For any real number d, the fractional integrated error process can be rewritten as

$$u_t = \sum_{k=0}^{\infty} \frac{\Gamma(k+d)}{\Gamma(d)\Gamma(k+1)} \varepsilon_{t-k}, \tag{5}$$

where $\Gamma(\cdot)$ is the gamma function. See Beran (1994, pp. 60). The long-run variance of u_t is one when $d = 0$, ∞ when $d > 1/2$, and zero when $d < 1/2$. Assuming $|\alpha| < 1$, we say that y_t is a long memory process whenever $d > 0$ as in Phillips and Shimotsu (2004) and a rough process when $d < 0$. The AR1FI model reduces to a standard autoregressive process when $d = 0$.

When $d \in (-1/2, 1/2)$, u_t is stationary and invertible (Bloomfield, 1985).[9] Let $\gamma_u(k) := Cov(u_t, u_{t-k})$ be the k^{th} order autocovariance of u_t. Under the specification of (4), according to Hosking (1981), the autocovariance function of u_t is

$$\gamma_u(k) = \frac{(-1)^k (-2d)!}{(k-d)!(-k-d)!} = \frac{(-1)^k \Gamma(1-2d)}{\Gamma(k-d+1)\Gamma(1-k-d)}, \tag{6}$$

where $(\cdot)!$ is the factorial of the argument. The k^{th} order ACF of u_t is

$$\rho_u(k) = \frac{(-d)!(k+d-1)!}{(d-1)!(k-d)!} \sim \frac{(-d)!}{(d-1)!} k^{2d-1} \text{ as } k \to \infty.$$

[7] https://dachxiu.chicagobooth.edu/#risklab.

[8] Despite its popularity in modeling the log RV, there are two limitations in the AR1FI(α, d) model. First, it fails to take account of estimation errors when the log RV is regarded as an estimator of the log quadratic variation. The estimation error necessitates a MA component; see Meddahi (2003) and Yu (2021). Second, it does not allow for jumps in the log volatility dynamic.

[9] The instantaneous variance of u_t is $E(u_t^2) = \frac{\Gamma(1-2d)}{(\Gamma(1-d))^2}$.

The correlation coefficient $\rho_u(k)$ decays at a hyperbolic rate as k goes to infinity. This is in contrast to the exponential decaying rate of an ARMA(p,q) model.

If $|\alpha| < 1$ and $d \in (-1/2, 1/2)$, y_t is covariance stationary and hence we can write $\gamma_y(k) := Cov(y_t, y_{t-k})$. Let $-\pi \leq \lambda \leq \pi$ be the Fourier frequency. The spectral density of y_t is

$$f_y(\lambda) = \frac{\sigma^2}{2\pi}\frac{(2-2\cos(\lambda))^{-d}}{1-2\alpha\cos(\lambda)+\alpha^2} \sim C\lambda^{-2d} \text{ when } \lambda \text{ is near zero,} \tag{7}$$

This is also the 'pseudo' spectral density (Velasco and Robinson, 2000) of AR1FI(α, d) when $d \in (1/2, 1)$.

4 Estimation Methods

In this section, we review four alternative estimation methods, namely, LWE, LWE(diff), the time-domain ML method, and the Whittle ML method.

4.1 LWE and LWE(diff)

Künsch (1987) and Robinson (1995a) investigate a class of models whose spectral densities satisfy the following property:

$$f_y(\lambda) \sim C\lambda^{-2d} \text{ as } \lambda \to 0^+ \tag{8}$$

with C being a positive constant. The property concerns only frequencies approaching zero.

The LWE method of Künsch (1987) and Robinson (1995a) is defined as

$$\begin{aligned}(\hat{C}, \hat{d}) &= \arg\max_{C,d} \frac{1}{m}\sum_{j=1}^{m}\left[-\log f_y\left(\lambda_j|\theta,\sigma_u^2\right) - \frac{I(\lambda_j)}{f_y(\lambda_j|\theta,\sigma_u^2)}\right] &(9)\\ &= \arg\max_{C,d} \frac{1}{m}\sum_{j=1}^{m}\left[-\log C + 2d\log\lambda_j - \frac{1}{C}\lambda^{2d} I(\lambda_j)\right], &(10)\end{aligned}$$

where $I(\lambda_j)$ denotes the periodogram at the j^{th} Fourier frequency $\lambda_j = 2\pi j/T$ with $j = 1, 2, \ldots, m$. Specifically,

$$I(\lambda_j) = \frac{1}{2\pi T}\left|\sum_{t=0}^{T} y_t \exp(-it\lambda_j)\right|^2, \tag{11}$$

which is a nonparametric estimate of the density. The parameter m satisfies the condition $m \leq (T-1)/2$ and diverges to infinity at a rate that is slower than T as $T \to \infty$. The analytical solution of LWE is

$$\hat{d} = \arg\max_{d}\left[-\log\hat{C}(d) + 2d\frac{1}{m}\sum_{j=1}^{m}\log\lambda_j\right] \text{ and } \hat{C}(d) = \frac{1}{m}\sum_{j=1}^{m}\lambda_j^{2d} I(\lambda_j). \tag{12}$$

Robinson (1995a) shows that the local Whittle estimator is consistent at the $\sqrt{m}$ rate and asymptotically normal with variance $1/(4m)$, that is,

$$\sqrt{m}\left(\hat{d}-d\right) \rightarrow_d N\left(0, 1/4\right),$$

when $d \in (-1/2, 1/2)$. Velasco (1999) investigates the possibility of using LWE for some non-stationary situations (i.e., $1/2 \leq d < 3/2$), showing that the consistency of LWE holds for $d \in (-1/2, 1)$ and the asymptotic normality holds for $d < 3/4$ with the same variance as in the stationary situation.[10]

There are significant advantages to using LWE. First, it works for a wider range of d, which goes beyond the stationary region $d \in (-1/2, 1/2)$. Second and perhaps most importantly, it is asymptotically robust against the short-run dynamic, which is determined by α in the AR1FI(α, d) model. However, the robustness comes with the cost of a reduced rate of convergence ($\sqrt{m}$ instead of $\sqrt{T}$). Moreover, a more significant and potentially empirically relevant problem is that LWE may have poor finite sample properties when the short-run dynamic is near unity. In this case, we expect $f_y(\lambda)$ is poorly approximated by $C\lambda^{-2d}$. At an intuitive level, the AR1FI($0, d$) model is observationally equivalent to the AR1FI($1, d-1$) model because

$$(1-L)\left(y_t-\mu\right) = \sigma_u \left(1-L\right)^{-d+1} \varepsilon_t$$

can be rewritten as

$$y_t - \mu = \sigma_u \left(1-L\right)^{-d} \varepsilon_t.$$

As a result, it is expected the spectral density of the AR1FI(α, d) model, whose α is strictly less than but very close to unity, is better approximated by $C\lambda^{-2d-2}$ when λ is close to zero. A detailed comparison between $\log(f_y(\lambda))$ and $\log(C\lambda^{-2d})$ will be made later in Figure 2.

When α is very close to unity, a sensible method to estimate d is to apply LWE to Δy_t, resulting in LWE(diff). If the estimated memory parameter by LWE(diff) is $\hat{d}$, within the class of AR1FI(α, d), it implies that the estimated model for y_t is either AR1FI($0, \hat{d}+1$) or AR1FI($1, \hat{d}$). Although LWE(diff) does not yield a consistent estimator of d when α is not exactly unity as discussed in Section 5.4, it may have good finite sample performances when α is very close to unity.

4.2 Time-domain ML Estimation

To implement the ML methods, we assume y_t is stationary, that is, $|\alpha| < 1$ and $d \in (-1/2, 1/2)$. The stationary assumption is imposed for two reasons. First, stationarity ensures that the likelihood function is relatively easier to calculate as elements in the variance-covariance matrix are finite and time-invariant. Second, most asset pricing models have been developed based on the condition that volatil-

[10] Shimotsu and Phillips (2006) propose an exact local Whittle estimation method, which can be applied to both stationary and non-stationary variables. Unlike the conventional local Whittle estimator, which approximates $I_u(\lambda_j)$ by $\lambda_j^{2d} I_y(\lambda_j)$, the exact local Whittle method is based on the relationship that $I_u(\lambda_j) = I_{\Delta^d y}(\lambda_j)$, where $I_{\Delta^d y}(\lambda_j)$ is the periodogram of $\Delta^d y = (1-L)^d y_t$. Shimotsu (2010) proposes a two-stage approach, which uses a tapered Local Whittle estimator (Velasco, 1999) in the first stage and a modified ELW objective function in the second stage. The 2-stage ELW method is designed to improve the performance of ELW when the mean (initial value) of the process is unknown. Unreported simulations show that both the ELW and the 2-stage ELW perform similar to the LWE method under our model setting.

ity is stationary. Examples include bond pricing (Duffie and Kan, 1996) and option pricing (Hull and White, 1987; Heston, 1993). See also the remark made by Robert Engle in Diebold (2003) against non-stationary volatility models.

Let $y = (y_1, y_2, \cdots, y_T)'$ and $\theta = (\alpha, d)$. Under the model specification of (3) with u_t specified as (4), $y_t - \mu$ follows a normal distribution with mean zero and variance-covariance matrix, denoted by Σ_y. The objective function of the ML estimator is given by

$$(\hat{\theta}, \hat{\sigma}_u) = \arg\max_{\theta, \sigma_u} \log L_N(\mu, \sigma_u, \theta),$$

where

$$\log L_N(\mu, \sigma_u, \theta) = \frac{1}{2T}\log|\Sigma_y| + \frac{1}{2T}(y - \mu l)' \Sigma_y^{-1}(y - \mu l), \tag{13}$$

and $l = (1, \ldots, 1)'$.

For the case of known mean value μ, the limiting properties of $\hat{\theta}$ was derived by Hannan (1973) for short memory processes and Yajima (1985) for long memory processes. That is, under some mild regularity conditions,

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \to_d N\left(0, \Xi_{\theta_0}^{-1}\right),$$

where θ_0 is the true parameter vector and Ξ_{θ_0} is the Fisher information matrix.

4.2.1 Modified profile likelihood

Dahlhaus (1989) extends the results of Yajima (1985) to the case with unknown mean. In case of unknown μ, a plug-in method is required. The plug-in method substitutes μ by a consistent estimator of the mean (e.g., the sample mean). Although the method provides a $\sqrt{T}$ consistent and asymptotically normal estimator, it is contaminated by an additional second-order negative bias (Lieberman, 2005) due to the need of estimating μ.

An alternative solution is the modified profile likelihood (MPL) estimator proposed by Cox and Reid (1987). The idea of the MPL estimator is to use a linear transformation of parameters of interest to make them orthogonal to nuisance parameters (μ and σ_u). The modified profile likelihood is given by

$$\log L_M(y, \hat{\mu}, \theta) = \left(\frac{1}{T} - \frac{1}{2}\right)\log|R| - \frac{1}{2}\log\left(l'R^{-1}l\right) + \frac{3-T}{2}\log\left[T^{-1}(y - \hat{\mu}l)' R^{-1}(y - \hat{\mu}l)\right], \tag{14}$$

where $R = \Sigma_y/\sigma_u^2$ and $\hat{\mu} = \left(l'R^{-1}l\right)^{-1} l'R^{-1}Y$. The asymptotic distribution of the MPL estimator is unchanged compared with the exact ML but eliminates some degree of bias in the exact ML (An and Bloomfield, 1993; Hauser, 1999).

4.2.2 Variance-covariance matrix Σ_y

Let the $(t,s)^{th}$ element of Σ_y be $\gamma_y(k)$, where $t,s=1,\cdots,T$ and $k=|t-s|$. The covariance function of the ARFIMA(p,d,q) process was derived by Hosking (1981, Lemma 1(c)) and Sowell (1992, eq. (8)-(9)) and approximated to improve computational speed by Chung (1994). In the special case of $p=1$ and $q=0$, the covariance function of Hosking (1981) is

$$\gamma_y(k)=\frac{\sigma_u^2}{1-\alpha^2}\gamma_u(k)\,A(k,\alpha)\,. \tag{15}$$

where $A(k,\alpha)=C(k,\alpha)+C(-k,\alpha)-1$, $C(k,\alpha)=F(d+k,1;1-d+k;\alpha)$, and $F(\cdot)$ is the hypergeometric function.

The hypergeometric function is computational costly and extremely large when k is large and α is far from unity,[11] leading to extreme behaviour of the covariance function. As an alternative, one may compute the covariance function as

$$\gamma_y(k)=\sigma_u^2\sum_{i=k}^{\infty}\sum_{j=0}^{\infty}\alpha^{i+j-k}\gamma_u(|i-j|)\,, \tag{16}$$

which resembles that of a standard AR(1) process. Unreported simulations show that the computed covariance values from (16) with a truncation of $20,000$ for both summands are identical to those obtained from (15) when the hypergeometric behaves normally (e.g., $\alpha=0.996$ and $k<1000$). This method is, however, quite computationally intensive.

Another method which is proposed by Bertelli and Caporin (2002) is referred to as the splitting approach. It is based on the following property of the covariance function for stationary processes (Brockwell and Davis, 2009):

$$\gamma_y(k)=\sum_{s=-\infty}^{\infty}\tilde{\gamma}(s)\,\gamma_u(k-s)\,, \tag{17}$$

where $\tilde{\gamma}(s)$ is the autocovariance of the pure AR component. For practical implementation, the summand is truncated at K.

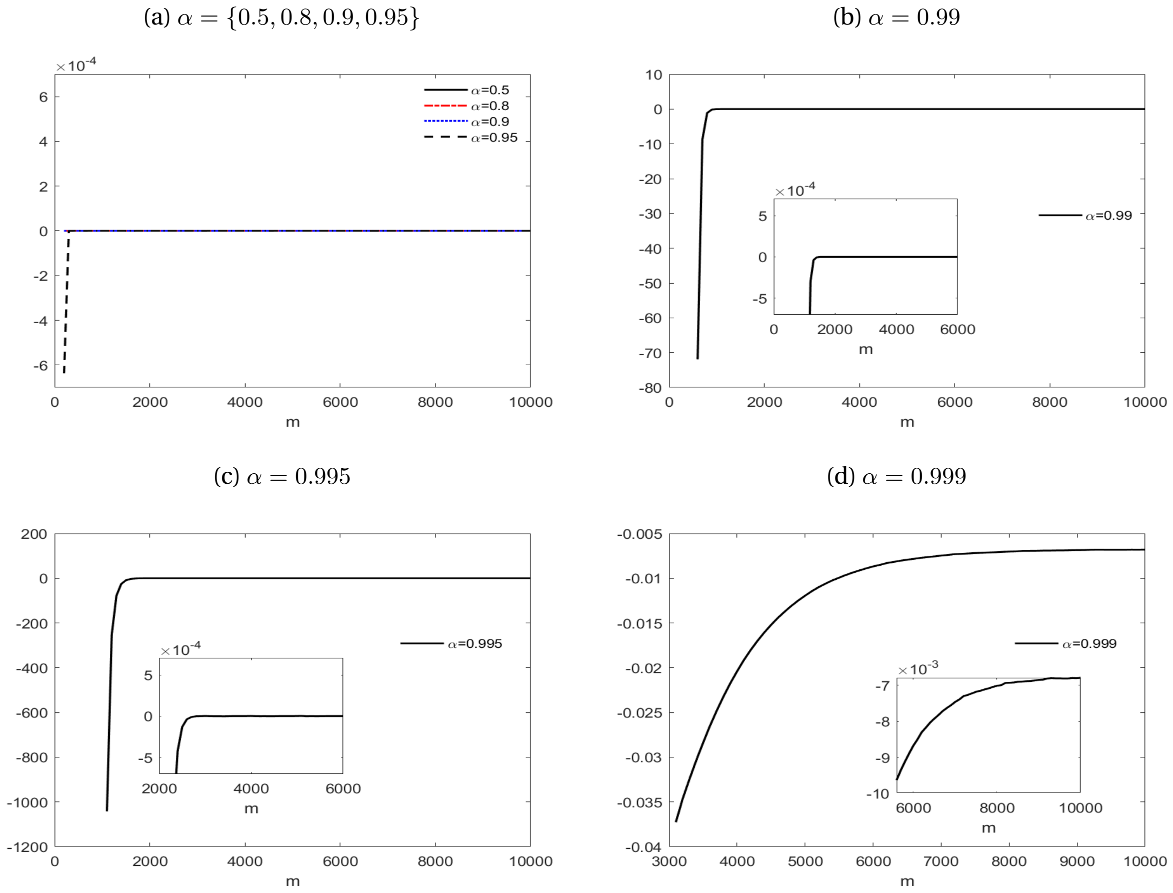
To provide some practical guidance for the choice of K, in Figure 1, we show the differences between the log determinants of Σ_y computed from (16) and (17) for each value of K, ranging from 1000 to 10000 (with an increment of 100). It is expected that one would need a larger K to ensure the estimation accuracy when the data series is highly persistent and has a long memory (i.e., when α is close to one and d is close to 0.5), as both γ_s and γ_u decay slower. We consider the autoregressive coefficient $\alpha=\{0.5,0.8,0.9,0.95,0.99,0.995,0.999\}$, $n=3000$ for the dimension of the covariance matrix, and $d=0.45$.[12] Evidently, the splitting method can provide very accurate estimation for the variance-covariance matrix with $K=200$ when $\alpha\leq 0.9$, which is consistent with the finding of Bertelli and Caporin (2002). However, for processes with autoregressive root close to unity, one would need a substantially larger value of K to ensure accuracy. Based on the simulation results presented in Figure

[11] For example, when $\alpha=0.93$, $d=0.4$ and $k=1,500$, $F(d+k,1;1-d+k;\alpha)=-1.2143\times 10^{19}$.

[12] Unreported simulations confirm that the estimation is more accurate than those presented here when $d<0.45$.

1, we recommend using $K = 300$ for $0.9 < \alpha \leq 0.95$, $K = 1700$ for $0.95 < \alpha \leq 0.99$, $K = 3000$ for $0.99 < \alpha \leq 0.995$, and $K = 7000$ for $0.995 < \alpha < 1$ so that the differences between the log determinants are at a maximum order of 10^{-3} when $d = 0.45$.

Figure 1: The differences between the log determinants of Σ_y computed from (16) and (17). The small plot within each subplot is a zoomed in version of the graph within a particular range.



(a) $\alpha = \{0.5, 0.8, 0.9, 0.95\}$

(b) $\alpha = 0.99$

(c) $\alpha = 0.995$

(d) $\alpha = 0.999$

4.3 Whittle ML Estimation

To avoid inverting Σ_y that is required in calculating the time-domain likelihood function, following Whittle (1953, 1954), one can approximate Σ_y^{-1} by $(2\pi)^{-2} \int_{-\pi}^{\pi} f_y(\lambda)^{-1} \cos((i-j)\lambda)\, d\lambda$ and $\log|\Sigma_y|$ by $T(2\pi)^{-1} \int_{-\pi}^{\pi} \log f_y(\lambda)\, d\lambda$ for a stationary process. The discrete-time version of the Whittle likelihood function (up to a scale multiplication) is

$$\log L_W\left(\theta, \sigma_u^2\right) = -\sum_{j=1}^{m} \log f_y\left(\lambda_j|\theta, \sigma_u^2\right) - \sum_{j=1}^{m} \frac{I(\lambda_j)}{f_y\left(\lambda_j|\theta, \sigma_u^2\right)}. \tag{18}$$

The Whittle likelihood function was presented in Künsch (1987) and Dahlhaus (1988). Fox and Taqqu (1986) show that the asymptotic properties of the estimators remain the same if we simplify the objective function to the following:

$$\log L_W\left(\theta,\sigma_u^2\right)' = -\sum_{j=1}^{m} \frac{I\left(\lambda_j\right)}{f_y\left(\lambda_j|\theta,\sigma_u^2\right)}, \tag{19}$$

where the distance between the spectrum density $f_y\left(\lambda_j|\theta,\sigma_u^2\right)$ and $I\left(\lambda_j\right)$ is minimized. We employ the simplified objective function for the estimation.[13] Like MPL, the parameter μ does not enter the objective function of the Whittle method as the zero frequency is not included. The spectral density of AR1FI(α, d) is given in (7). The Whittle ML method yields $\sqrt{T}$-consistent, asymptotically normal and efficient parameter estimates (Hannan, 1973; Fox and Taqqu, 1986; Giraitis and Surgailis, 1990) when $d \in (0, 1/2)$.

5 Monte Carlo Simulations

We now examine the finite sample properties of various estimation techniques. The DGP is AR1FI(α, d),[14] covering both Model 1 and Model 2. We assume α takes a value in {-0.2, 0, 0.3, 0.5, 0.7, 0.9, 0.99, 0.996} and d takes a value in {-0.4, 0, 0.4}. We set $\sigma_u = 1$ and μ to zero but assume them unknown. The initial value of each simulated sample path is set to the long-run mean (i.e., $\mu/(1-\alpha)$), which is zero under this setting. The first $5,000$ observations are discarded from each simulated sample-path to minimize the impact of the initial value. The number of replications is $1,000$.

We investigate the estimation accuracy of the semi-parametric methods and the ML methods for both the memory parameter d and the short-run dynamic parameter α. Table 1 provides a brief summary of the existing literature on the simulation and their Monte Carlo designs. Our Monte Carlo design extends those in the existing studies by considering more empirically relevant parameter values. In particular, we (1) allow maximum value of α to be much closer to the unity (i.e. 0.996 versus 0.8 in Smith, Taylor, and Yadav (1997) and Nielsen and Frederiksen (2005) and 0.9 in Nadarajah, Martin, and Poskitt (2021)); (2) consider larger sample sizes (i.e. $T = 1024$ as well as $T = 2048, 4096$ in the case $\alpha \geq 0.9$ for LWE). The choices of near unit α and sample size are guided by the empirical results that will be reported later.

Table 1: Existing Monte Carlo Studies

Paper	Relevant Tables	Relevant Estimation Methods	Sample Size
Smith, Taylor, and Yadav (1997)	Tables I and VI	ML and LPE ($m = T^{0.5}$, $T^{0.6}$,$T^{0.7}$)	256
Nielsen and Frederiksen (2005)	Tables 8 and 9	Exact ML, MPL, Whittle, Conditional ML LWE and LPE ($m = T^{0.5}$, $T^{0.65}$)	$128, 256, 512$
Nadarajah, Martin, and Poskitt (2021)	Tables 6 and 7	ML and LPE ($m = T^{0.65}$)	$96, 576$

Implementing the time-domain ML method under these parameter settings is not straightforward, as existing methods for computing the variance-covariance matrix do not work. As discussed in Sec-

[13] Coursol and Dacunha-Castelle (1982) study the approximation error $\log L_N - \log L_W$.

[14] The fractionally integrated process in (4) is simulated with the *fracdiff* function provided by Katsumi Shimotsu.

tion 4.2.2, the hypergeometric function in the formulas of Hosking (1981), Sowell (1992), and Chung (1994) behaves abnormally when the dimension of the matrix (which is the same as T) is large. The suggested truncation of $K = 200$ in (17) of the splitting method cannot provide accurate results when α is close to unity. We propose an alternative truncation scheme as detailed in Section 4.2.2. For the semi-parametric methods, in addition to the popular LWE and LPE methods, we also investigate the finite sample performance of LWE(diff). The LWE(diff) method has already been used in empirical applications (e.g., Phillips and Shimotsu (2004)), but its finite sample performance has yet been studied.

Several interesting findings emerge from the simulations. In particular, we document the poor finite sample performance of LWE (LWE(diff)) when the short-run dynamic is strong (weak) and explain why. We show that the ML estimators have a bi-modal distribution under certain parameter settings, leading to a possible mis-identification between Model 1 and 2. Under this circumstance, the traditional performance measures such as mean and standard deviations are not appropriate. Alternative measures are used to present the estimation results. A summary of the simulation findings is provided in Section 5.3.

5.1 Semiparametric Methods

We first investigate the performance of LWE and LWE(diff). The parameter support for d is $(-1, 3/2)$ and the bandwidth $m = \{\lfloor T^{0.55}\rfloor, \lfloor T^{0.65}\rfloor, \lfloor T^{0.75}\rfloor, \lfloor T^{0.85}\rfloor\}$, where $\lfloor . \rfloor$ denotes the integer part of the argument. The objective functions are optimized with the command *fminbnd* in MATLAB, as there is only one model parameter. Table 2 reports the mean and standard error (in brackets) of the LWE and LWE(diff) estimates of $\hat{d}$, obtained from 1000 replications. The sample size is set at $T = 1024$.[15] There are several interesting observations from Table 2.

First, LWE works very well in estimating d when α is near zero (say $\alpha \leq 0.3$), with negligible biases and small standard errors. This is especially true when $m = \lfloor T^{0.65}\rfloor$. Together with its asymptotic robustness property against short-run dynamics, the good finite sample property may be the reason why it has been popular in estimating d for log RVs. However, it leads to a substantial upward bias (spurious long memory) when the process becomes more persistent. The substantial upward bias in d by the semi-parametric methods when $\alpha = 0.8$ or 0.9 and $T = \{96, 256, 512, 576\}$ has been documented in Smith et al. (1997), Nielsen and Frederiksen (2005), and Nadarajah et al. (2021). Our results indicate that this upward bias problem continues to hold when $\alpha = \{0.99, 0.996\}$ and $T = 1028$. Interestingly, the bias increases towards one as α gets closer to unity. For example, when α is 0.996 and d is -0.4 (i.e. Model 2 is the DGP), with a small standard error of 0.06, the estimated d (with $m = \lfloor T^{0.65}\rfloor$) is located around 0.58, always suggesting spurious long memory. This is expected because, when α is very close to unity, the spectral density is better approximated by $C\lambda^{-2d-2}$ and hence, LWE essentially estimates $d + 1$.

To better understand this point, we show the gaps between the theoretical spectral density of y_t and

[15] We set the sample size to be the power of two to ensure the accuracy of Fourier transformation. Moreover, the finite sample properties reported here remain qualitatively unchanged when T is increased to 2408 and 4096. The results may be obtained from the authors upon request.

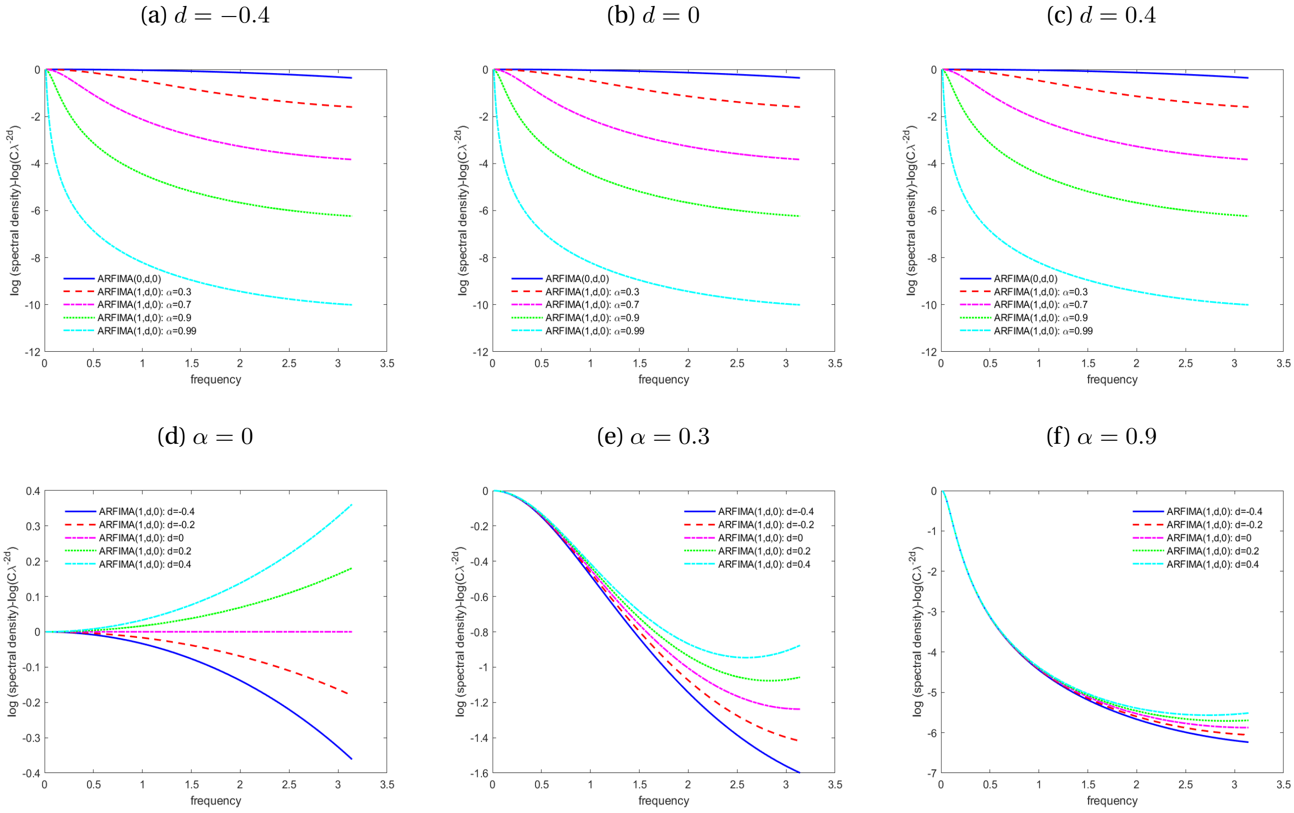
Table 2: Mean and standard error (in bracket) of LWE and LWE(diff) of d with $m = \{T^{0.55}, T^{0.65}, T^{0.75}, T^{0.85}\}$ and $T = 1024$

	LWE				LWE(diff)			
	$m = T^{0.55}$	$m = T^{0.65}$	$m = T^{0.75}$	$m = T^{0.85}$	$m = T^{0.55}$	$m = T^{0.65}$	$m = T^{0.75}$	$m = T^{0.85}$
$\alpha = -0.2$								
d=-0.4	-0.39 (0.10)	-0.40 (0.06)	-0.41 (0.04)	-0.45 (0.03)	-0.49 (0.26)	-0.68 (0.21)	-0.82 (0.15)	-0.92 (0.09)
d=0	-0.01 (0.09)	-0.01 (0.06)	-0.03 (0.04)	-0.07 (0.03)	-0.64 (0.21)	-0.74 (0.16)	-0.82 (0.12)	-0.88 (0.09)
d=0.4	0.40 (0.09)	0.39 (0.06)	0.38 (0.04)	0.31 (0.03)	-0.55 (0.11)	-0.57 (0.07)	-0.59 (0.05)	-0.63 (0.03)
$\alpha = 0$								
d=-0.4	-0.39 (0.10)	-0.40 (0.06)	-0.39 (0.04)	-0.38 (0.03)	-0.56 (0.25)	-0.73 (0.19)	-0.85 (0.13)	-0.92 (0.09)
d=0	-0.01 (0.09)	-0.01 (0.06)	-0.00 (0.04)	-0.00 (0.03)	-0.69 (0.20)	-0.77 (0.15)	-0.83 (0.11)	-0.84 (0.08)
d=0.4	0.40 (0.09)	0.40 (0.06)	0.40 (0.04)	0.38 (0.03)	-0.56 (0.10)	-0.57 (0.07)	-0.57 (0.05)	-0.56 (0.03)
$\alpha = 0.3$								
d=-0.4	-0.39 (0.09)	-0.37 (0.06)	-0.32 (0.04)	-0.21 (0.03)	-0.66 (0.23)	-0.80 (0.16)	-0.88 (0.12)	-0.90 (0.10)
d=0	-0.00 (0.09)	0.02 (0.06)	0.07 (0.04)	0.17 (0.03)	-0.74 (0.17)	-0.80 (0.13)	-0.81 (0.09)	-0.73 (0.05)
d=0.4	0.41 (0.09)	0.42 (0.06)	0.47 (0.04)	0.55 (0.03)	-0.56 (0.10)	-0.56 (0.07)	-0.51 (0.04)	-0.39 (0.03)
$\alpha = 0.5$								
d=-0.4	-0.38 (0.09)	-0.33 (0.06)	-0.22 (0.04)	-0.04 (0.03)	-0.73 (0.21)	-0.84 (0.14)	-0.88 (0.11)	-0.85 (0.09)
d=0	0.01 (0.09)	0.06 (0.06)	0.17 (0.04)	0.33 (0.03)	-0.77 (0.15)	-0.80 (0.11)	-0.74 (0.07)	-0.58 (0.04)
d=0.4	0.42 (0.09)	0.46 (0.06)	0.57 (0.04)	0.71 (0.03)	-0.55 (0.10)	-0.52 (0.06)	-0.41 (0.04)	-0.23 (0.03)
$\alpha = 0.7$								
d=-0.4	-0.33 (0.09)	-0.21 (0.06)	-0.02 (0.04)	0.18 (0.03)	-0.80 (0.18)	-0.86 (0.13)	-0.84 (0.11)	-0.71 (0.05)
d=0	0.06 (0.09)	0.18 (0.06)	0.37 (0.04)	0.55 (0.03)	-0.78 (0.13)	-0.73 (0.09)	-0.58 (0.05)	-0.38 (0.03)
d=0.4	0.47 (0.09)	0.58 (0.06)	0.77 (0.04)	0.92 (0.03)	-0.51 (0.10)	-0.41 (0.06)	-0.21 (0.04)	-0.01 (0.03)
$\alpha = 0.9$								
d=-0.4	-0.04 (0.10)	0.17 (0.07)	0.34 (0.05)	0.43 (0.03)	-0.80 (0.15)	-0.72 (0.09)	-0.60 (0.05)	-0.50 (0.03)
d=0	0.36 (0.09)	0.57 (0.07)	0.74 (0.05)	0.80 (0.03)	-0.58 (0.10)	-0.40 (0.07)	-0.24 (0.05)	-0.13 (0.03)
d=0.4	0.76 (0.10)	0.96 (0.07)	1.12 (0.05)	1.16 (0.04)	-0.23 (0.10)	-0.02 (0.07)	0.15 (0.05)	0.24 (0.03)
$\alpha = 0.99$								
d=-0.4	0.52 (0.09)	0.55 (0.06)	0.56 (0.04)	0.56 (0.03)	-0.45 (0.10)	-0.43 (0.06)	-0.41 (0.04)	-0.39 (0.03)
d=0	0.91 (0.09)	0.94 (0.06)	0.96 (0.04)	0.93 (0.03)	-0.08 (0.09)	-0.05 (0.06)	-0.03 (0.04)	-0.01 (0.03)
d=0.4	1.21 (0.11)	1.21 (0.11)	1.20 (0.11)	1.14 (0.11)	0.32 (0.09)	0.35 (0.06)	0.37 (0.04)	0.36 (0.03)
$\alpha = 0.996$								
d=-0.4	0.57 (0.09)	0.58 (0.06)	0.58 (0.04)	0.57 (0.03)	-0.41 (0.09)	-0.41 (0.06)	-0.40 (0.04)	-0.38 (0.03)
d=0	0.97 (0.09)	0.97 (0.06)	0.97 (0.04)	0.94 (0.03)	-0.03 (0.09)	-0.02 (0.06)	-0.01 (0.04)	-0.01 (0.03)
d=0.4	1.19 (0.13)	1.17 (0.12)	1.14 (0.11)	1.08 (0.11)	0.37 (0.09)	0.38 (0.06)	0.38 (0.04)	0.38 (0.03)

the approximate spectral density $C\lambda^{-2d}$ used by LWE under various parameter settings. The larger the distance between $f(\lambda)$ and $C\lambda^{-2d}$ is, the less accurate estimated results are expected from LWE. Figure 2 plots the quantity $\log(f(\lambda)) - \log(C\lambda^{-2d})$ against the frequency λ. We choose the value of C such that the quantity takes value zero at frequency zero. It is obvious that the distances at frequencies close to zero are affected substantially by α but not so much by d. This is consistent with our findings in Table 2 that LWE leads a substantial bias when α is close to unity, while the bias is similar across various values of d given a value of α.

Second, there is a trade-off between bias and standard error with the different choice of m for LWE. When the bandwidth parameter m reduces from $\lfloor T^{0.85} \rfloor$ to $\lfloor T^{0.55} \rfloor$, the bias of the d estimate decreases. However, the use of smaller tuning parameter m does not alleviate the problem of severe bias in LWE of d when α is very close to unity (i.e., $\alpha = 0.996$).

Figure 2: The difference between the theoretical spectral density and the approximate spectral density for AR1FI(α, d): $\log(f(\lambda)) - \log(C\lambda^{-2d})$


(a) $d = -0.4$

(b) $d = 0$

(c) $d = 0.4$

(d) $\alpha = 0$

(e) $\alpha = 0.3$

(f) $\alpha = 0.9$

The finite sample problem in LWE for the case when α is close to unity naturally suggests that one may use LWE(diff) to estimate d. The right panel of Table 2 reports the mean and standard error of LWE(diff) under the same parameter settings. There are several interesting observations. First, as expected, the performance of LWE(diff) is good when $\alpha = 0.996$. The mean and standard error are nearly a mirror image of those of LWE when $\alpha = 0$. In general, LWE(diff) works very well in estimating d when α is near unity (say $\alpha \geq 0.99$), with negligible biases and small standard errors. The results are relatively stable across different settings of m, with $m = \lfloor T^{0.85} \rfloor$ providing estimates with the smallest variations. Second, LWE(diff) leads to a substantial downward bias in d when α is not so close to zero, including the case $\alpha = 0.9$. The further α away from unity, the larger the downward bias is. For example, when the true value of d is 0.4 and the true value of α is 0.3 (i.e., the true DGP is Model 1), with a small standard error of 0.03, the estimated d (with $m = T^{0.85}$) is located around -0.39, always suggesting spurious anti-persistent errors.

To understand if larger sample sizes can help address the finite sample problems in LWE, in Table 3 we report the mean and standard error of LWE with $m = \lfloor T^{0.65} \rfloor$, when $T = \{2048, 4096\}$ and $\alpha = \{0.9.0.99, 0.996\}$, obtained from 1000 replications. The sample sizes $T = \{2048, 4096\}$ are large but remain empirically reasonable for RV. For the ease of comparison, we also report the results when $T = 1024$. It is clear that while the standard error reduces as T increases, the bias remains substantial.

For example, when α is 0.996 and d is -0.4 (i.e. Model 2 is the DGP) and $T = 4096$, with a small standard error of 0.04, the estimated d is located around 0.57, always suggesting spurious long memory. Similar finite sample problems apply to LWE(diff) when $T = \{2048, 4096\}$ and α is far away from unity.

Table 3: Mean and standard error (in bracket) of LWE of d with $m = \lfloor T^{0.65} \rfloor$ and $T = 1024, 2048, 4096$.

	$\alpha = 0.9$	$\alpha = 0.99$	$\alpha = 0.996$
$T = 1024$			
d=-0.4	0.17 (0.07)	0.55 (0.06)	0.58 (0.06)
d=0	0.57 (0.07)	0.94 (0.06)	0.97 (0.06)
d=0.4	0.96 (0.07)	1.21 (0.11)	1.17 (0.12)
$T = 2048$			
d=-0.4	0.09 (0.05)	0.54 (0.05)	0.57 (0.05)
d=0	0.49 (0.05)	0.93 (0.05)	0.97 (0.05)
d=0.4	0.89 (0.05)	1.24 (0.09)	1.21 (0.11)
$T = 4096$			
d=-0.4	0.00 (0.04)	0.52 (0.04)	0.57 (0.04)
d=0	0.40 (0.04)	0.92 (0.04)	0.96 (0.04)
d=0.4	0.80 (0.04)	1.26 (0.07)	1.24 (0.10)

To obtain an estimate of α using LWE, we fit an AR(1) model to pre-filtered data series using $\hat{d}$ obtained from LWE.[16] This two-stage approach has been used in the literature; see, for example, Andersen, Bollerslev, Diebold, and Labys (2003). The last column of Table 5 reports the mean and standard error of $\hat{\alpha}$, based on LWE ($m = \lfloor T^{0.65} \rfloor$), for the same parameter setting as before and $T = 1024$. From Table 5, the estimated α from LWE is fairly close to its true value when $\alpha \leq 0.3$. However, when $\alpha > 0.3$, the upward biases in $\hat{d}$ lead to equally significant downward biases in $\hat{\alpha}_1$. When $d = -0.4$ and $\alpha = \{0.99, 0.996\}$ (i.e. the true DGP is Model 2), with a small standard error, LWE tends to conclude that α is located around 0. Together with the simulation results on $\hat{d}$ reported earlier, we conclude that LWE always suggests that the estimated model is Model 1 when the DGP is Model 2. Once again, this finding is not surprising as AR1FI(α, d) with $\alpha = 0.99, 0.996$ is very similar to AR1FI$(0, d+1)$.

Our simulation studies suggest that one should be cautious against using LWE and LWE(diff). LWE tends to point to Model 1 when the DGP is Model 2; LWE(diff) tends to point to Model 2 when the DGP is Model 1. Since we do not know the value of α *ex ante* in practice, we generally do not know if we should use LWE or LWE(diff).

5.2 Parametric Methods

For the Whittle method, we use a grid searching method to choose the 'optimal' initial values of d and α. The grids range from -0.499 to 0.499 for d and from -0.999 to 0.999 for α, with an increment of 0.005. We evaluate the Whittle log-likelihood for all possible combinations of d and α. The pair that produces the highest log-likelihood value is taken as our initial values for the Whittle method. For

[16] There is no need to estimate α using LWE(diff) as it assumes $\alpha = 1$.

MPL, we set the initial values of the two parameters to be the estimates of the Whittle method. For both MPL and Whittle, the parameter supports for α and d are $(-1, 1)$ and $(-0.5, 0.5)$, respectively. The log-likelihoods of the two parametric ML methods are optimized using the *fmincon* function in MATLAB with the sequential quadratic programming algorithm. The two ML methods estimate both d and α simultaneously.

Following the common practice in the literature, in Tables 4-5, we report the means and standard errors of $\hat{\alpha}$ and $\hat{d}$ for MPL, Whittle, and Whittle (taper) under the same parameter settings as before with $T = 1028$, obtained from all 1000 replications.[17] For the ease of comparison, we repeat results of LWE and LWE(diff) in the last two columns.

Table 4: Mean and standard error (in bracket) of $\hat{d}$ when $T = 1024$. The bandwidth parameter $m = \lfloor T^{0.65} \rfloor$ for LWE and $m = \lfloor T^{0.85} \rfloor$ for LWE(diff). Boldface corresponds to cases where the DGP is Model 1 or Model 2.

	MPL	Whittle	Whittle (taper)	LWE	LWE(diff)
$\alpha = -0.2$					
d=-0.4	-0.40 (0.04)	-0.40 (0.04)	-0.41 (0.04)	-0.40 (0.06)	-0.97 (0.17)
d=0	-0.00 (0.04)	-0.01 (0.04)	-0.01 (0.04)	-0.01 (0.06)	-0.88 (0.10)
d=0.4	0.40 (0.04)	0.39 (0.04)	0.40 (0.04)	0.39 (0.06)	-0.63 (0.03)
$\alpha = 0$					
d=-0.4	-0.40 (0.05)	-0.41 (0.04)	-0.42 (0.05)	-0.40 (0.06)	-0.97 (0.15)
d=0	-0.01 (0.04)	-0.01 (0.04)	-0.02 (0.05)	-0.01 (0.06)	-0.84 (0.08)
d=0.4	**0.29 (0.28)**	**0.29 (0.28)**	**0.31 (0.26)**	0.40 (0.06)	-0.56 (0.03)
$\alpha = 0.3$					
d=-0.4	-0.41 (0.06)	-0.41 (0.06)	-0.42 (0.06)	-0.37 (0.06)	-0.92 (0.12)
d=0	-0.01 (0.08)	-0.03 (0.08)	-0.04 (0.11)	0.02 (0.06)	-0.73 (0.05)
d=0.4	**0.35 (0.17)**	**0.33 (0.17)**	**0.34 (0.18)**	0.42 (0.06)	-0.39 (0.03)
$\alpha = 0.5$					
d=-0.4	-0.41 (0.07)	-0.42 (0.07)	-0.43 (0.07)	-0.33 (0.06)	-0.85 (0.09)
d=0	-0.03 (0.10)	-0.06 (0.11)	-0.07 (0.13)	0.06 (0.06)	-0.58 (0.04)
d=0.4	0.38 (0.10)	0.35 (0.11)	0.36 (0.12)	0.46 (0.06)	-0.23 (0.03)
$\alpha = 0.7$					
d=-0.4	-0.39 (0.08)	-0.42 (0.08)	-0.42 (0.08)	-0.21 (0.06)	-0.71 (0.05)
d=0	-0.00 (0.09)	-0.04 (0.09)	-0.03 (0.10)	0.18 (0.06)	-0.38 (0.03)
d=0.4	0.39 (0.08)	0.36 (0.08)	0.38 (0.09)	0.58 (0.06)	-0.01 (0.03)
$\alpha = 0.9$					
d=-0.4	-0.39 (0.06)	-0.40 (0.06)	-0.40 (0.07)	0.17 (0.07)	-0.50 (0.03)
d=0	0.01 (0.06)	-0.00 (0.05)	0.00 (0.07)	0.57 (0.07)	-0.13 (0.03)
d=0.4	0.41 (0.05)	0.36 (0.06)	0.40 (0.06)	0.96 (0.07)	0.24 (0.03)
$\alpha = 0.99$					
d=-0.4	**-0.38 (0.13)**	**-0.38 (0.13)**	**-0.37 (0.15)**	0.55 (0.06)	-0.39 (0.03)
d=0	0.00 (0.03)	0.00 (0.03)	0.00 (0.03)	0.94 (0.06)	-0.01 (0.03)
d=0.4	0.43 (0.05)	0.19 (0.12)	0.41 (0.03)	1.21 (0.11)	0.36 (0.03)
$\alpha = 0.996$					
d=-0.4	**-0.38 (0.12)**	**-0.38 (0.12)**	**-0.36 (0.18)**	0.58 (0.06)	-0.38 (0.03)
d=0	0.00 (0.03)	0.00 (0.03)	0.01 (0.03)	0.97 (0.06)	-0.01 (0.03)
d=0.4	0.44 (0.05)	0.13 (0.11)	0.42 (0.03)	1.17 (0.12)	0.38 (0.03)

[17] Tapering has been shown capable of removing deterministic time trends (e.g., Žurbenko (1979); Robinson (1986); Dahlhaus (1988); Hurvich and Ray (1995); Velasco (1999); Hurvich and Chen (2000)). The tapering methods are detailed in the Appendix.

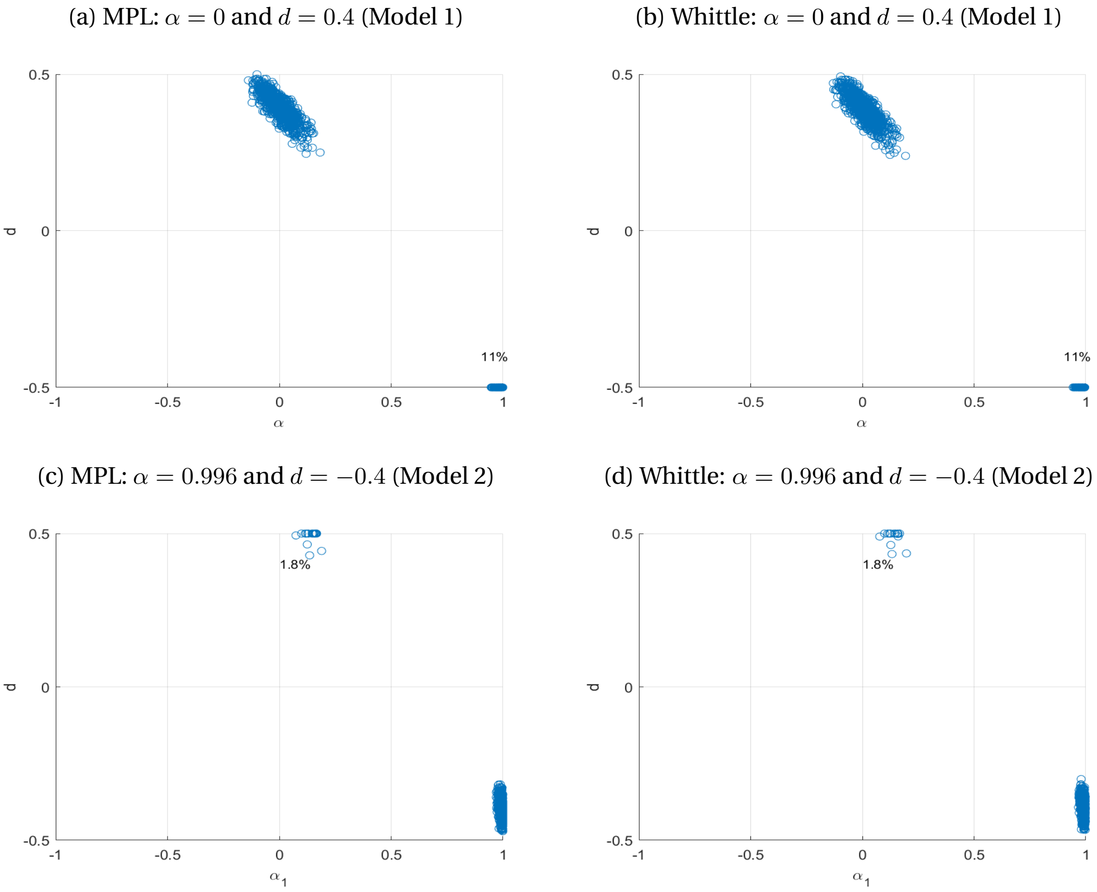
Table 5: Mean and standard error (in bracket) of $\hat{\alpha}$ when $T = 1024$. The bandwidth parameter $m = \lfloor T^{0.65} \rfloor$ for LWE. Boldface corresponds to cases where the DGP is Model 1 or Model 2.

	MPL	Whittle	Whittle (taper)	LWE
		$\alpha = -0.2$		
d=-0.4	-0.20 (0.05)	-0.20 (0.04)	-0.19 (0.05)	-0.20 (0.06)
d=0	-0.20 (0.04)	-0.19 (0.04)	-0.19 (0.05)	-0.19 (0.06)
d=0.4	-0.20 (0.04)	-0.20 (0.04)	-0.20 (0.05)	-0.19 (0.06)
		$\alpha = 0$		
d=-0.4	0.00 (0.06)	0.01 (0.05)	0.01 (0.06)	-0.00 (0.07)
d=0	0.00 (0.05)	0.01 (0.05)	0.01 (0.06)	0.01 (0.07)
d=0.4	**0.11 (0.31)**	**0.11 (0.31)**	**0.09 (0.28)**	0.00 (0.07)
		$\alpha = 0.3$		
d=-0.4	0.30 (0.07)	0.31 (0.07)	0.32 (0.07)	0.27 (0.07)
d=0	0.31 (0.09)	0.33 (0.09)	0.34 (0.12)	0.28 (0.07)
d=0.4	**0.35 (0.17)**	**0.36 (0.17)**	**0.36 (0.17)**	0.28 (0.07)
		$\alpha = 0.5$		
d=-0.4	0.50 (0.09)	0.52 (0.07)	0.52 (0.08)	0.43 (0.07)
d=0	0.52 (0.10)	0.55 (0.11)	0.56 (0.12)	0.44 (0.06)
d=0.4	0.52 (0.10)	0.55 (0.11)	0.53 (0.11)	0.43 (0.06)
		$\alpha = 0.7$		
d=-0.4	0.69 (0.08)	0.71 (0.07)	0.71 (0.08)	0.52 (0.06)
d=0	0.69 (0.08)	0.72 (0.07)	0.72 (0.09)	0.53 (0.06)
d=0.4	0.70 (0.07)	0.72 (0.07)	0.71 (0.08)	0.53 (0.06)
		$\alpha = 0.9$		
d=-0.4	0.89 (0.05)	0.89 (0.04)	0.89 (0.05)	0.37 (0.08)
d=0	0.89 (0.04)	0.89 (0.04)	0.89 (0.05)	0.37 (0.08)
d=0.4	0.89 (0.03)	0.91 (0.03)	0.90 (0.03)	0.39 (0.08)
		$\alpha = 0.99$		
d=-0.4	**0.97 (0.12)**	**0.97 (0.12)**	**0.96 (0.14)**	0.05 (0.07)
d=0	0.99 (0.01)	0.98 (0.01)	0.98 (0.01)	0.06 (0.08)
d=0.4	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	-0.03 (0.46)
		$\alpha = 0.996$		
d=-0.4	**0.98 (0.11)**	**0.97 (0.11)**	**0.95 (0.17)**	0.02 (0.07)
d=0	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.03 (0.09)
d=0.4	0.99 (0.01)	0.99 (0.00)	0.99 (0.00)	-0.01 (0.56)

As evidenced in the two tables, although the two ML methods generally work well across all parameter settings for both d and α, there are two important exceptions that are empirically relevant to RV. When the true DGP is Model 1 or Model 2 (highlighted in boldface in the two tables), the standard errors are unusually large. For example, from Table 4, if $\alpha = 0$, the standard error of $\hat{d}$ for both MPL and Whittle is 0.28 when $d = 0.4$, which is about six times larger than those when $d = 0$; if $\alpha = 0.99$, the standard errors of $\hat{d}$ for MPL and Whittle (taper) are 0.13 and 0.15 when $d = -0.4$, which are about four times larger than those when $d = 0$. Similar features are observed for $\hat{\alpha}$ from Table 5.

These unusually large standard errors motivate us to examine the finite sample property in different ways. In Figure 3 we report the scatter plots of the estimated d and α by MPL and Whittle from the 1000 replications when the true parameter values are $\alpha = 0$ and $d = 0.4$ (Model 1) and $\alpha = 0.996$ and $d = -0.4$ (Model 2). The scatter plots indicate that when the true DGP is Model 1 or Model 2, two disjoint clusters are obtained, one located around the true parameter values and the other one is far away. Interestingly, one of the clusters corresponds to Model 1 and the other corresponds to Model 2. That is, the finite sample distribution is a mixture of two disjoint distributions. In the figures, we also report the percentage of replications (out of 1000 replications) that fall in each cluster. When the true parameter values are $\alpha = 0$ and $d = 0.4$, with probability 89% (or for 890 replications), the two ML methods yield estimates around the true values; with probability 11%, the two ML methods yield an estimate of α near unity and an estimate of d near -0.5. When the true parameter values are $\alpha = 0.996$ and $d = -0.4$, with probability 98.2%, the two ML methods yield estimates around the true values; with probability 1.8% (or for 18 replications), the two ML methods yield an estimate of α near zero and an estimate of d near 0.5.

Figure 3: Scatter plots of the estimated d and α by MPL and Whittle from the 1000 simulated paths when $\alpha = 0$ and $d = 0.4$. The number on the graph is the percentage of replications where the estimates fall in the wrong parameter region.



(a) MPL: $\alpha = 0$ and $d = 0.4$ (Model 1)

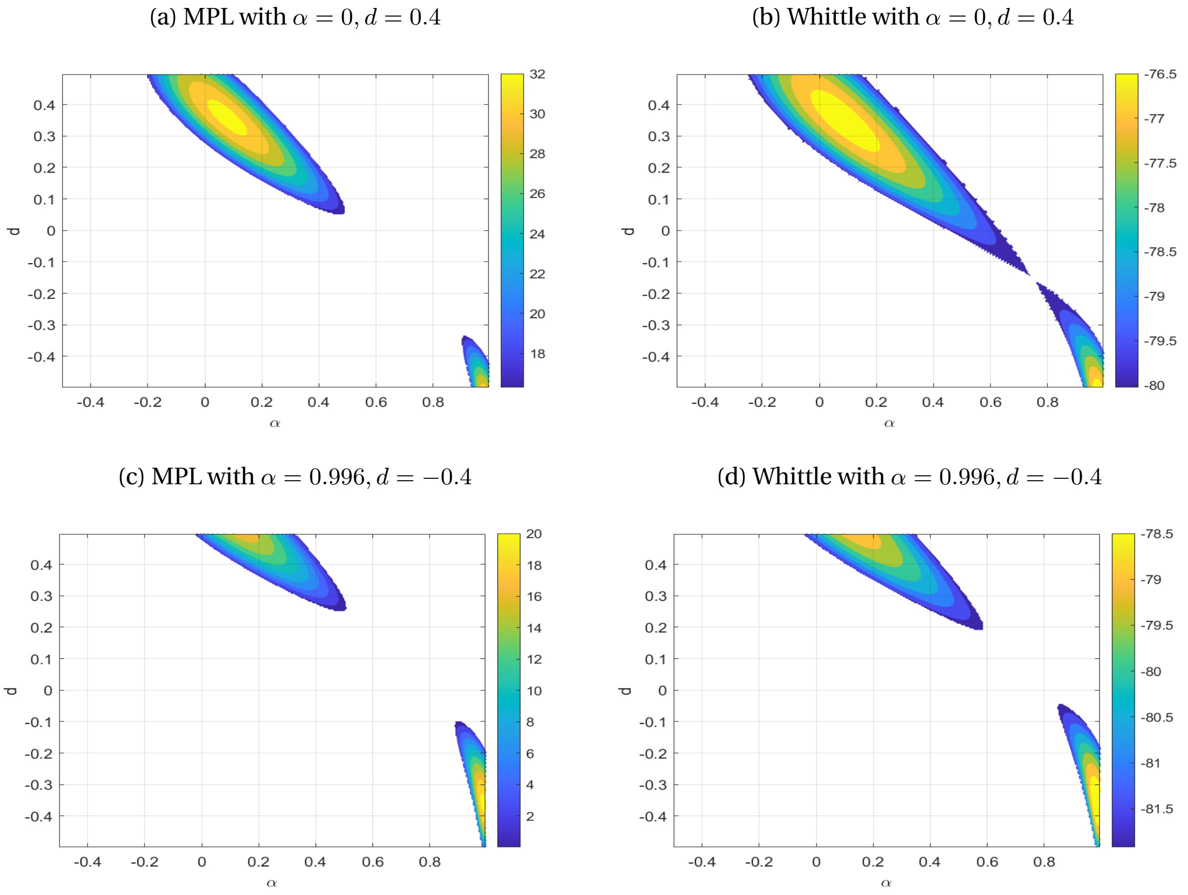
(b) Whittle: $\alpha = 0$ and $d = 0.4$ (Model 1)

(c) MPL: $\alpha = 0.996$ and $d = -0.4$ (Model 2)

(d) Whittle: $\alpha = 0.996$ and $d = -0.4$ (Model 2)

The scatter plots indicate that it is very plausible that there are two modes in the likelihood functions for MPL and Whittle. Figure 4 displays the contour plots of the log-likelihood surfaces of MPL (left panels) and Whittle (right panels) when the data is generated from $\alpha = 0$ and $d = 0.4$ (top panels) or from $\alpha = 0.996$ and $d = -0.4$ (bottom panels). We remove log-likelihood values that are smaller than certain thresholds to obtain better visualization of the surface at the peak. The two modes can be seen in all cases.

Figure 4: Contour plots of the log likelihood surfaces of MPL and Whittle for a simulated sample path under the settings of $\alpha = 0$ and $d = 0.4$ (top panels) and $\alpha = 0.996$ and $d = -0.4$ (bottom panels).



(a) MPL with $\alpha = 0, d = 0.4$

(b) Whittle with $\alpha = 0, d = 0.4$

(c) MPL with $\alpha = 0.996, d = -0.4$

(d) Whittle with $\alpha = 0.996, d = -0.4$

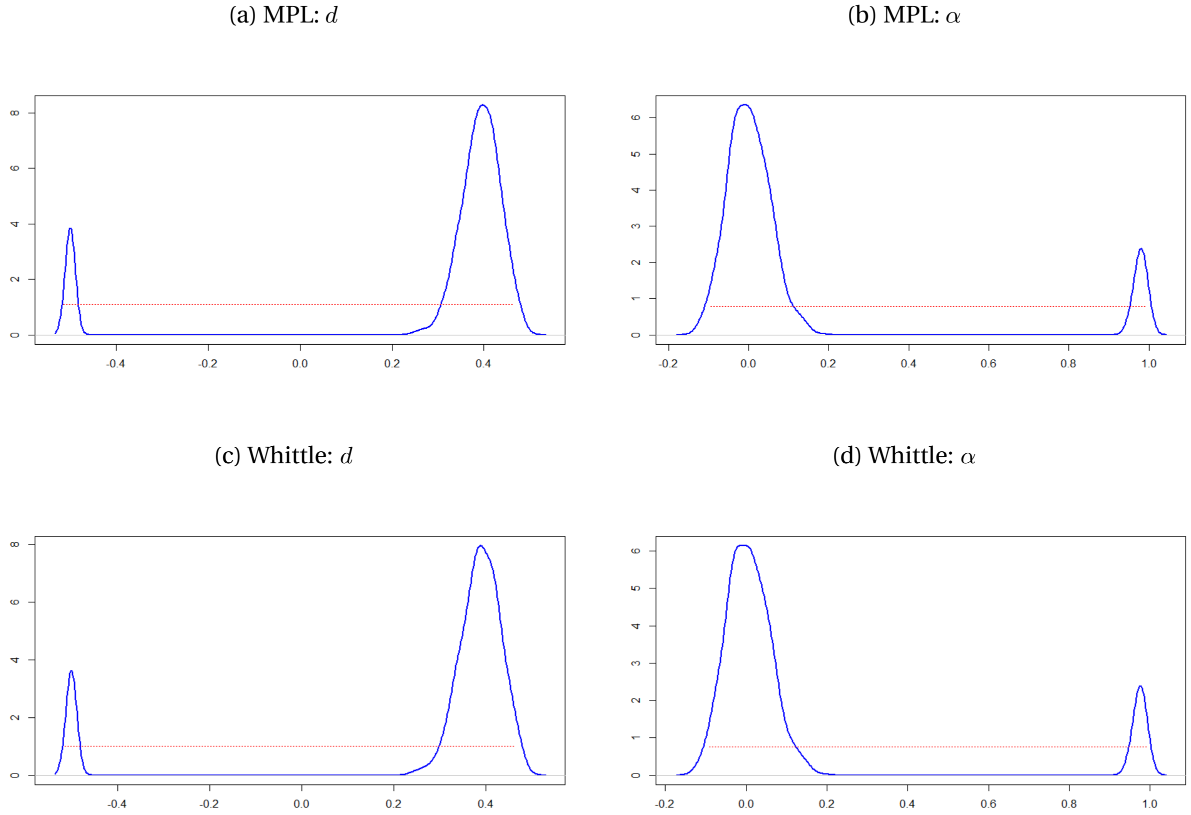
When the distribution is bi-modal, mean and standard error are not ideal performance measures. In Table 6, for each of the two parametric ML methods, we report the mean and standard error of $\hat{d}$ and $\hat{\alpha}$ for each cluster (instead of the whole distribution) and the probability of false identification (PFI) when the DGP is from Model 1 (i.e., $(d, \alpha) = (0.4, 0)$ or $(0.4, 0.3)$) and from Model 2 (i.e., $(d, \alpha) = (-0.4, 0.99)$ or $(-0.4, 0.996)$). For the ease of comparison, we also report the mean and standard errors for the whole distribution (all 1000 replications). It can be seen that when α is close to zero, the mean and standard errors of the correct cluster for both ML methods compare well with those of LWE. When α is close to unity, the mean and standard errors of the correct cluster for both ML methods compare well with those of LWE(diff). Moreover, when $d = 0.4$ and $\alpha = 0$, the standard error of $\hat{d}$ for the two ML methods, obtained from all 1000 replications, is larger than that in other cases. This is because PFI is the largest (11%) in this case.

Table 6: Mean and standard error (in bracket) of $\hat{d}$ and $\hat{\alpha}$ of each cluster and all replications, and the probability of false identification (PFI). In all cases $T = 1024$.

	$(d, \alpha) = (0.4, 0)$	$(d, \alpha) = (0.4, 0.3)$	$(d, \alpha) = (-0.4, 0.99)$	$(d, \alpha) = (-0.4, 0.996)$
Whittle	[0.29 (0.28),0.11 (0.31)]	[0.33 (0.17),0.36 (0.17)]	[-0.38 (0.13),0.97 (0.12)]	[-0.38 (0.12),0.97 (0.11)]
Whittle (correct)	[0.39 (0.04),0.01 (0.05)]	[0.38 (0.06),0.32 (0.07)]	[-0.40 (0.03),0.98 (0.01)]	[-0.39 (0.03),0.99 (0.01)]
Whittle (false)	[-0.50 (0.00),0.97 (0.01)]	[-0.29 (0.06),0.94 (0.04)]	[0.47 (0.03),0.15 (0.03)]	[0.49 (0.02),0.14 (0.03)]
Whittle: PFI	0.110	0.062	0.022	0.018
MPL	[0.29 (0.28),0.11 (0.31)]	[0.35 (0.17),0.35 (0.17)]	[-0.38 (0.13),0.97 (0.12)]	[-0.38 (0.12),0.98 (0.11)]
MPL (correct)	[0.39 (0.04),0.00 (0.05)]	[0.39 (0.06),0.31 (0.07)]	[-0.40 (0.03),0.99 (0.01)]	[-0.40 (0.03),0.99 (0.01)]
MPL (false)	[-0.50 (0.00),0.98 (0.01)]	[-0.29 (0.07),0.94 (0.04)]	[0.47 (0.03),0.15 (0.03)]	[0.49 (0.02),0.14 (0.03)]
MPL: PFI	0.110	0.061	0.022	0.018

In Figure 5 we plot the kernel densities of $\hat{d}$ and $\hat{\alpha}$ (solid lines) for MPL and Whittle when $d = 0.4$ and $\alpha = 0$ from 1000 replications. The bi-modality in the finite sample distribution can be seen clearly. In the same figure, we also show the 95% highest density intervals (the shortest confidence intervals), identified by the two segments of the dotted line around each mode. Not surprisingly, in all cases the highest density interval contains two disjoint intervals.

Figure 5: The kernel densities (solid lines) of $\hat{d}$ and $\hat{\alpha}$ for MPL and Whittle when $d = 0.4, \alpha = 0$. The two segments of the dotted line around each mode form the 95% highest density interval.


(a) MPL: d

(b) MPL: α

(c) Whittle: d

(d) Whittle: α

5.3 Summary of Monte Carlo Studies

Our simulation results suggest that the LWE and LWE(diff) methods show significant bias when the autoregressive coefficient of the process is not close to zero and unity, respectively. Most importantly, the semi-parametric methods cannot separate Model 1 and 2. While LWE can provide accurate estimation results under Model 1, it always falsely points to Model 1 when the true DGP is Model 2, even with large sample sizes. This finding explains why research papers employing semi-parametric methods (LWE or LPE) tend to find evidence of long memory with d around 0.5 (Model 1). The LWE(diff) is the opposite of LWE. It works well under Model 2 but fails to yield a satisfactory estimation outcome for d when α is far away from unity.

The ML methods work well in general. However, interestingly and new to the literature, we found that when the true DGP is either Model 1 or Model 2, the log-likelihood surfaces of the two ML methods have two modes. One mode corresponds to Model 1, and the other corresponds to Model 2. Although the mode around the true parameter values is generally higher than the other mode, the other way around is possible. That is why the ML methods cannot separate Model 1 and 2. Thus, not surprisingly, research papers employing the two parametric ML methods have found both long memory and roughness in log volatilities. Suppose one finds Model 1 (Model 2) by the ML methods in empirical applications. It is very likely that the true model is Model 1 (Model 2), but there is also a nonnegligible probability that the true DGP is Model 2 (Model 1).

5.4 Discussions: Model 1 versus Model 2

Model 1 can be formulated as a fractional process with a local-to-zero AR coefficient such that

$$\text{Model 1}: y_t = -\frac{c}{T}y_{t-1} + \sigma\left(1-L\right)^{-\tilde{d}}\varepsilon_t \text{ with constant } c \text{ and } \tilde{d} > 0. \tag{20}$$

Model 2 can be understood as a local-to-unity AR(1) model with fractionally integrated errors. That is,

$$\text{Model 2}: y_t = \left(1-\frac{c}{T}\right)y_{t-1} + \sigma\left(1-L\right)^{-d}\varepsilon_t \text{ with } c \geq 0 \text{ and } d \in \left(-0.5, 0\right), \tag{21}$$

which can be rewritten as

$$\Delta y_t = -\frac{c}{T}y_{t-1} + \sigma\left(1-L\right)^{-d}\varepsilon_t. \tag{22}$$

When $c = 0$, Model 1 and 2 become AR1FI$(0, \tilde{d})$ and AR1FI$(1, d)$, respectively. When $\tilde{d} = d + 1$, these two models are observationally equivalent. Hence, Model 1 and 2 can be viewed as local alternatives of these two observationally equivalent models, respectively, with the same local deviation quantity (i.e., $-\frac{c}{T}y_{t-1}$). That is why Model 1 and 2 can generate similar sample paths and it is difficult to distinguish them in finite samples when the local deviation (c) is small.

Although Model 1 and 2 deviate from the two observationally equivalent models by the same quantity, they have different asymptotic properties, unless $c = 0$. Multiplying both sides of Model 1 by

$(1-L)$ leads to

$$\Delta y_t = -\frac{c}{T}\Delta y_{t-1} + \sigma\left(1-L\right)^{-d}\varepsilon_t.$$

As $T \to \infty$, the first term on the right-hand side is dominated by the second term. The process is asymptotically equivalent to AR1FI$(1,d)$ — a model employed by LWE(diff). From Theorem 2 of Davydov (1970), as $T \to \infty$,

$$\frac{\delta_d \Gamma\left(1+d\right)}{T^{d+0.5}\sigma} y_{\lfloor Tr \rfloor} \Rightarrow B_H\left(r\right), \tag{23}$$

where $B_H\left(r\right)$ is the fBm with Hurst parameter $H = d + 0.5 \in (0, 0.5)$ and $r \in [0,1]$. The fBm process is used in Gatheral, Jaisson, and Rosenbaum (2018) to model and forecast log RV where $H = 0.14$ is assumed.

For Model 2, when $c > 0$, from Tanaka (2013), as $T \to \infty$,

$$\frac{\delta_d \Gamma\left(1+d\right)}{T^{d+0.5}\sigma} y_{\lfloor Tr \rfloor} \Rightarrow J_c^H\left(r\right), \tag{24}$$

where $\delta_d = \sqrt{\frac{2(d+0.5)\Gamma(1-d)}{\Gamma(1+d)\Gamma(1-2d)}}$ and $J_c^H\left(r\right) := \exp\left(cr\right)\int_0^r \exp\left(-cs\right) dB_H\left(s\right)$ is a fractional OU process. Equation (21) can be rewritten as

$$y_t = -\frac{c}{T}\left(1-L\right)^{-1} y_{t-1} + \sigma\left(1-L\right)^{-\tilde{d}}\varepsilon_t. \tag{25}$$

The first term on the right-hand side of (25) represents the difference between Model 2 with $\alpha < 1$ and the LWE(diff) model $(\alpha = 1)$ and has the following limiting:

$$-\frac{\delta_d \Gamma\left(1+d\right)}{T^{d+0.5}\sigma}\frac{c}{T}\left(1-L\right)^{-1} y_{t-1} \Rightarrow J_c^H\left(r\right) - B_H\left(r\right).$$

Since $-\frac{c}{T}\left(1-L\right)^{-1} y_{t-1} = O_p\left(T^{d+0.5}\right)$ which is of the same order of magnitude as the second quantity in (22), one cannot ignore this term even asymptotically whenever $c \neq 0$. As shown in the forecasting exercise in the application section, it matters more for long horizon forecasting.

6 Empirical Applications to RV

In this section, we investigate the dynamics of log RVs for the S&P 500 index exchange traded fund (ETF) and the nine industry ETFs over the past decade. The sample period starts from 5 January 2010 and ends on 25 May 2021. The QML realized volatility data are obtained from the Risk Lab. Assets under consideration are listed in Table 7, along with the number of observations and summary statistics of log RV (QMLE) of each data series.

Table 7: Summary statistics of the log realized volatility (QMLE) of various financial assets

Ticker	Obs.	Mean	Std.	Skew.	Kurto.
S&P 500 ETF (SPY)	2757	-2.40	0.50	0.66	4.04
Consumer discretionary (XLY)	2722	-2.22	0.46	0.73	4.01
Consumer staples (XLP)	2723	-2.42	0.41	1.28	7.31
Energy (XLE)	2724	-1.81	0.44	0.78	3.92
Financial (XLF)	2724	-2.04	0.43	0.95	5.31
Health Care (XLV)	2723	-2.25	0.41	0.98	5.27
Industrial (XLI)	2724	-2.18	0.45	0.75	4.32
Material (XLB)	2723	-2.06	0.44	0.64	4.14
Technology (XLK)	2723	-2.18	0.46	0.78	4.65
Utilities (XLU)	2723	-2.09	0.37	1.04	6.82

6.1 Estimation Results

The AR1FI(α, d) model is fitted to each (demeaned) log RV series using the two semi-parametric and two parametric methods. The bandwidth $m = \lfloor T^{0.65} \rfloor$ for LWE and $m = \lfloor T^{0.85} \rfloor$ for LWE(diff). For the Whittle method, we use the same grid searching method as in the simulation studies. For MPL, we use the estimation results of the Whittle method as the initial value.

The estimated parameters are reported in Table 8. The two ML methods provide almost identical results for all assets. The estimated autoregressive coefficients are close to unity and the estimated fractional parameters are close to those of LWE(diff). These estimates point to Model 2 and are consistent with the findings of Gatheral, Jaisson, and Rosenbaum (2018); Wang, Xiao, and Yu (2021), where the fBm ad fOU processes are fitted to log RV. They are also consistent with those of Liu, Shi, and Yu (2020); Fukasawa, Takabatake, and Westphal (2021); Bolko, Christensen, Pakkanen, and Veliyev (2021), where log spot volatility is treated as latent but assumed to follow an AR(1) process with fractionally integrated errors or fBm.

Table 8: Model estimation results

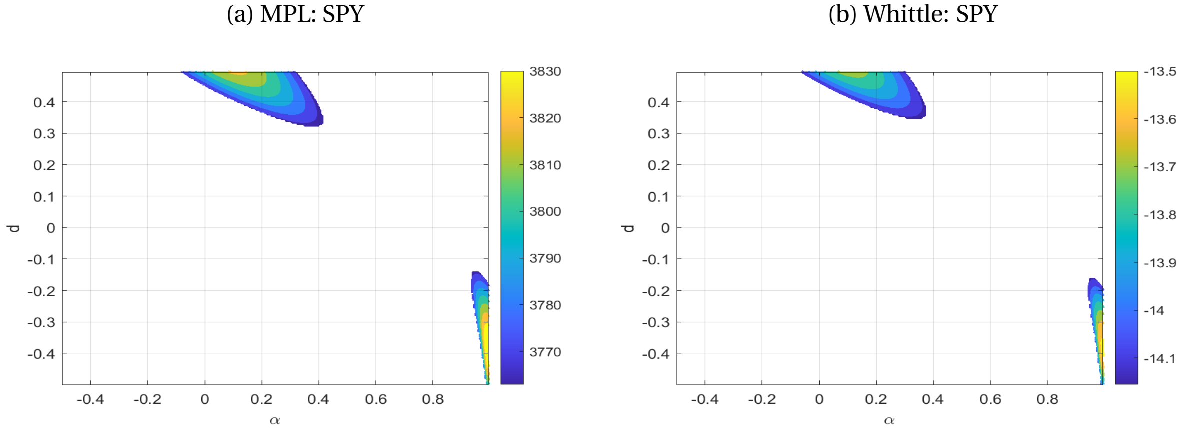
Name (Ticker)	MPL		Whittle		LWE		LWE(diff)
	$\hat{d}$	$\hat{\alpha}$	$\hat{d}$	$\hat{\alpha}$	$\hat{d}$	$\hat{\alpha}$	$\hat{d}$
S&P 500 ETF (SPY)	-0.382	0.995	-0.382	0.995	0.602	-0.006	-0.356
Consumer discretionary (XLY)	-0.446	0.999	-0.445	0.998	0.643	-0.105	-0.413
Consumer staples (XLP)	-0.387	0.998	-0.385	0.997	0.698	-0.093	-0.373
Energy (XLE)	-0.427	0.998	-0.427	0.997	0.607	-0.053	-0.396
Financial (XLF)	-0.422	0.998	-0.423	0.997	0.611	-0.056	-0.390
Health Care (XLV)	-0.440	0.998	-0.439	0.997	0.540	0.004	-0.412
Industrial (XLI)	-0.468	0.997	-0.467	0.996	0.582	-0.110	-0.410
Material (XLB)	-0.470	0.998	-0.469	0.997	0.632	-0.162	-0.415
Technology (XLK)	-0.426	0.995	-0.426	0.994	0.601	-0.061	-0.397
Utilities (XLU)	-0.434	0.998	-0.434	0.997	0.607	-0.070	-0.400

On the contrary, LWE suggests that the memory parameter is between 0.54 and 0.70, implying that all the log RV series have a long memory. The autoregressive coefficient is always close to zero, suggesting Model 1. The estimated fractional parameters from LWE(diff) are between -0.41 and -0.36. The difference between the two estimates of d by LWE and LWE(diff) is almost one, as expected. The results from LWE(diff) are close to those of the ML methods.

The seemingly contradictory results between LWE and other methods are consistent with our simulation results. Suppose the true DGP is Model 2 as the ML methods suggested. According to our simulations, LWE will lead to the false conclusion of Model 1. LWE(diff) performs well under this setting and hence, leads to similar outcomes as the parametric methods. Suppose the estimation results of the ML methods are incorrect and the true DGP is Model 1. Our simulations suggest that LWE performs well in this case, and employing LWE(diff) will lead to a false conclusion of Model 2.

Based on the simulation studies, the two ML methods are relatively more reliable than LWE. As such, it is more likely that the log RVs of the ETFs are generated from Model 2 than from Model 1. However, there is still a nonnegligible probability of false identification as the log-likelihood surfaces have two modes with similar values. As an illustration, we show in Figure 6 contour plots of the log-likelihood surfaces of MPL and Whittle for the log RV of SPY. Again, we remove log-likelihood values that are far from the peak to enable better visibility of the modes. Evidently, there are two modes in the likelihood surface. One is around Model 1, and the other one is around Model 2. The color around Model 2 is brighter, implying higher likelihood values in the region and hence the estimation outcome.

Figure 6: Contour plots of the log likelihood surfaces of MPL and Whittle for SPY.



(a) MPL: SPY

(b) Whittle: SPY

6.2 Model Forecasting

The one-step-ahead linear prediction of AR1FI(α, d) can be written as

$$\hat{y}_{t+1} = \phi_{t,1}y_t + \phi_{t,2}y_{t-1} + \cdots + \phi_{t,t}y_1 \text{ with } t > 1. \tag{26}$$

We employ the popular Durbin-Levinson algorithm for the computation of the forecasting coefficients $\phi_{t,j}$ with $j = 1, 2, \cdots, t$. Specifically, under the assumption that $\{y_t\}$ is a zero mean stationary process with autocovariance function $\gamma_y(.)$ such that $\gamma_y(0) > 0$ and $\gamma_y(k) \to 0$ as $k \to \infty$, $\phi_{1,1} = \gamma_y(1)/\gamma_y(0)$, $v_0 = \gamma_y(0)$,

$$\begin{aligned} v_t &= v_{t-1}\left[1 - \phi_{t,t}^2\right] \text{ for } t = 1, 2, \ldots, \\ \phi_{t,t} &= \left[\gamma_y(t) - \sum_{j=1}^{t-1} \phi_{t-1,j}\gamma_y(t-j)\right] v_{t-1}^{-1}, \\ \begin{bmatrix} \phi_{t,1} \\ \vdots \\ \phi_{t,t-1} \end{bmatrix} &= \begin{bmatrix} \phi_{t-1,1} \\ \vdots \\ \phi_{t-1,t-1} \end{bmatrix} - \phi_{t,t} \begin{bmatrix} \phi_{t-1,t-1} \\ \vdots \\ \phi_{t-1,1} \end{bmatrix}. \end{aligned}$$

See Brockwell and Davis (1987, chp. 5). The two-step-ahead recursive forecasting is computed recursively as

$$\hat{y}_{t+2} = \phi_{t+1,1}\hat{y}_{t+1} + \phi_{t+1,2}y_t + \cdots + \phi_{t+1,t+1}y_1$$

and the k-step-ahead forecast can be obtained in a similar fashion.

The computation of $\gamma_y(t-j)$ (and hence $\phi_{t,j}$) of AR1FI(α, d) requires three model parameters: d, α, σ^2. The σ^2 parameter is, however, not estimated directly. Here, we estimate it by

$$\hat{\sigma}^2 = \frac{1}{T}\sum_t e_t^2 \text{ with } e_t = (1-L)^{\hat{d}}(y_t - \hat{\alpha}_1 y_{t-1}).$$

We conduct the forecasting exercise on the log RV (QML) of the ten financial assets. The out-of-sample forecasting period starts from January 5, 2018. The model parameters are estimated from a rolling window with size of eight years. The forecasting horizon ranges from one period to 50 periods.

Forecasting Evaluation

The second step is to evaluate the forecasting accuracy. We consider two loss functions: mean squared forecast error (MSFE) and mean absolute forecast error (MAFE). They are defined as

$$\begin{aligned} MSFE_k &= \frac{1}{(T-T_0+1)}\sum_{t=T_0+1}^{T}(\hat{y}_{t+k} - y_{t+k})^2, \\ MAFE_k &= \frac{1}{(T-T_0+1)}\sum_{t=T_0+1}^{T}|\hat{y}_{t+k} - y_{t+k}|, \end{aligned}$$

where T_0 is the total number of observations in the training sample and T is the total sample size.

To assess whether the competing models are statistically different, we employ the model confidence set (MCS) approach proposed by Hansen et al. (2011). The approach aims to provide a model confidence set which contains the best models with probability greater than or equal to a pre-specified level, say 5%. The MCS is expected to be large when the data does not contain sufficient information to tell the models apart. It also provides 'p-value' for each individual model.

Let the set of competing models be $\mathcal{M}_0$ with objects indexed by $i = 1, \ldots, M$. We have $M = 4$ in our setting. The loss function, denoted by $L_{i,t}$, can either be the squared error $\left(\hat{y}^{(i)}_{t+k} - y_{t+k}\right)^2$ or the absolute error $\left|\hat{y}^{(i)}_{t+k} - y_{t+k}\right|$, where $\hat{y}^{(i)}_{t+k}$ is the k-step-ahead forecast from model i. The relative performance is measured by $d_{i,j,t} = L_{i,t} - L_{j,t}$ for all $i, j \in \mathcal{M}_0$. The MCS procedure applies tests of

$$H_{0,\mathcal{M}} : E\left(d_{ij,t}\right) = 0 \text{ for all } i, j \in \mathcal{M} \subset \mathcal{M}_0,$$

against the alternative

$$H_{A,\mathcal{M}} : E\left(d_{ij,t}\right) \neq 0 \text{ for some } i, j \in \mathcal{M}.$$

The procedure is based on a model equivalence test for the null $H_{0,\mathcal{M}}$ for any $\mathcal{M} \subset \mathcal{M}_0$ and an elimination rule that identifies the object to be removed from $\mathcal{M}$ if $H_{0,\mathcal{M}}$ is rejected. The algorithm is implemented as follows. The initial model set is $\mathcal{M} = \mathcal{M}_0$.

Step 1 Test $H_{0,\mathcal{M}}$ using a model equivalence test at level α.

Step 2 If $H_{0,\mathcal{M}}$ is 'accepted' set $\hat{\mathcal{M}}_{1-\alpha} = \mathcal{M}$, otherwise use an elimination rule to remove objects from $\mathcal{M}$ and repeat Step 1 and 2.

The set $\hat{\mathcal{M}}_{1-\alpha}$ is referred to as the model confidence set. Let $\mathcal{M}_i$ be the model set tested in the i^{th} iteration, $P_{H_{0,\mathcal{M}_i}}$ be the p-value associated with the null hypothesis $H_{0,\mathcal{M}_i}$, and $e_{\mathcal{M}_i}$ be the element to be eliminated from set $\mathcal{M}_i$ in the event that $H_{0,\mathcal{M}_i}$ is rejected. The MCS p-value for model $e_{\mathcal{M}_i}$ is defined by

$$\hat{p}_{e_{\mathcal{M}_i}} = \max_{j \leq i} P_{H_{0,\mathcal{M}_j}},$$

where $\mathcal{M}_1 \supset \mathcal{M}_2 \ldots \supset \mathcal{M}_i$.

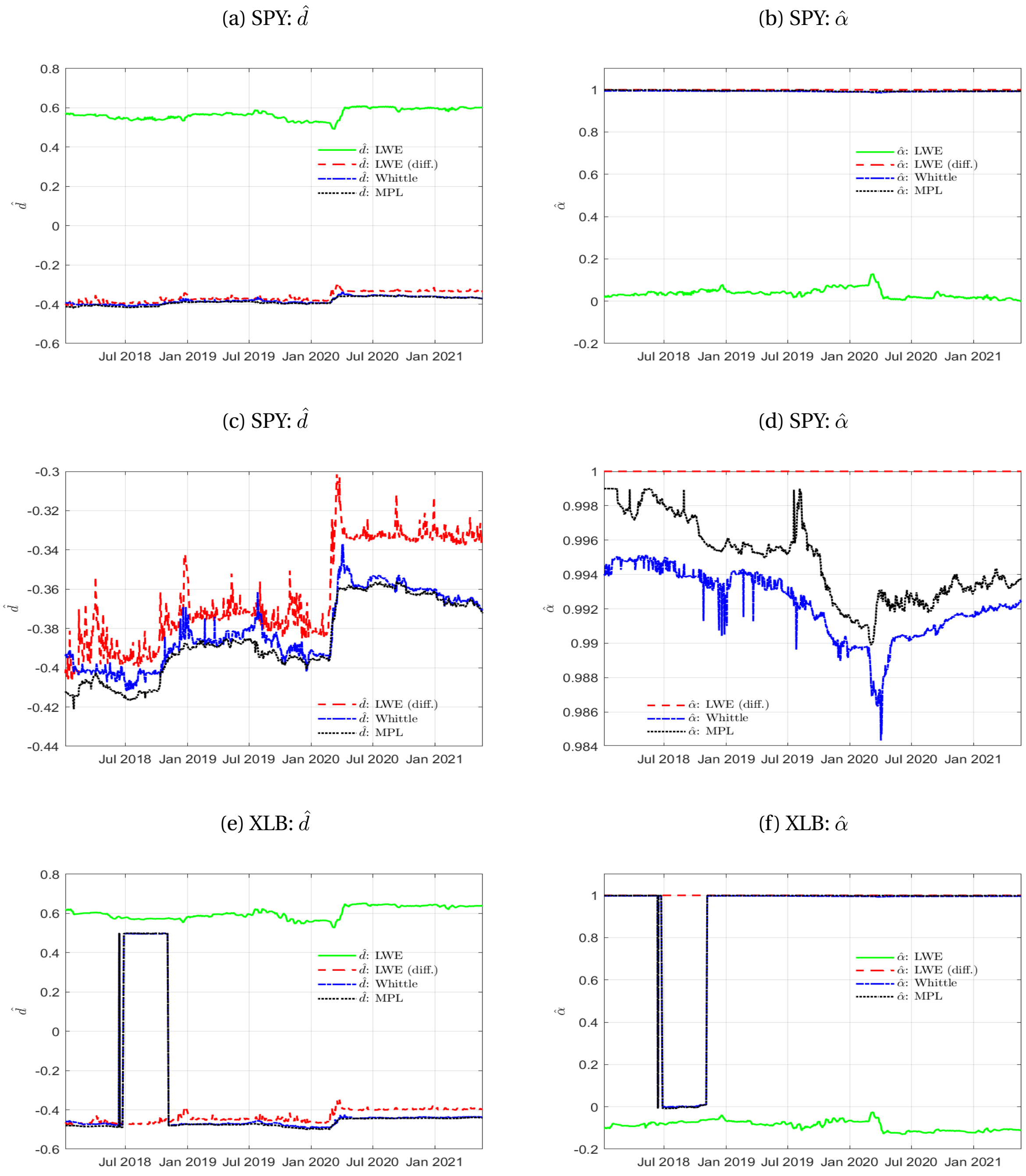
For the model equivalence test, we employed the $T_{max,\mathcal{M}}$ statistic with a bootstrapped implementation as recommended by Hansen et al. (2011). The block length of the bootstraps is set to be 20. See Hansen et al. (2011) for details of the test statistic and its associated elimination rule.

Forecasting Results

We compare the out-of-sample performance of the four estimation methods in forecasting the log RVs. The general conclusion for all of the data series are the same. That is, the Whittle method provides the best out-of-sample forecasting results (especially in the long run), followed by MPL.

For the ease of presentation, we take the S&P500 ETF (SPY) and the material industry ETF (XLB) as two examples. We employ a rolling window algorithm for the out-of-sample forecast, where the model is re-estimated for each subsample. Figure 7 displays the rolling window estimate of d and α from SPY and XLB. The top row shows estimation results from all four methods and the second row displays only those from LWE(diff), Whittle, and MPL for SPY. Consistent with the whole sample analysis, results from LWE(diff), Whittle, and MPL are close to each other, with the estimated d fluctuating around -0.4 and the estimated α being either one or very close to one. While the LWE method suggests that α is close to zero and d is greater than 0.5. Furthermore, from the second row of the graph, we can see that the Whittle estimates of d seem to be between those of MPL and LWE(diff) for most of the rolling windows, while the Whittle estimates of α are consistently the lowest among the three methods.

Figure 7: The rolling window estimates of the model parameters



(a) SPY: $\hat{d}$

(b) SPY: $\hat{\alpha}$

(c) SPY: $\hat{d}$

(d) SPY: $\hat{\alpha}$

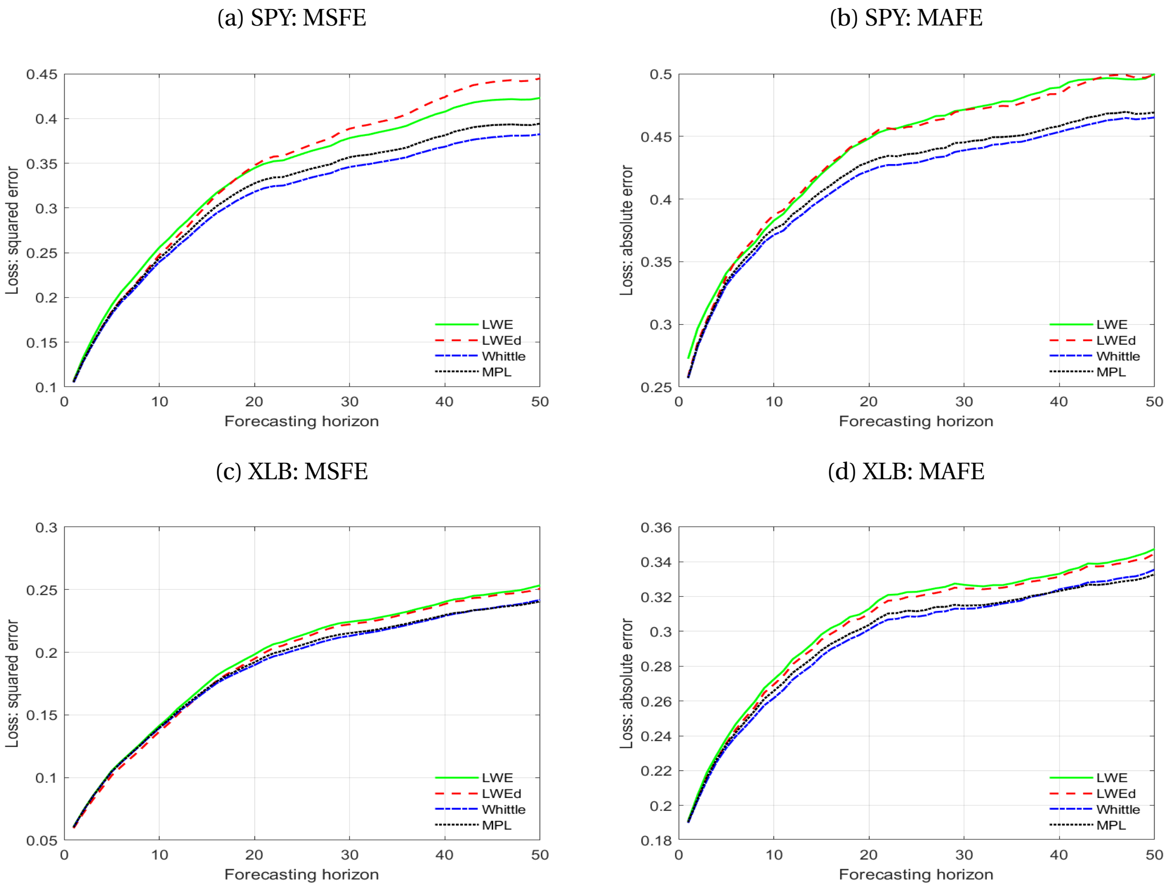
(e) XLB: $\hat{d}$

(f) XLB: $\hat{\alpha}$

The rolling window estimates of d and α from XLB are displayed in the bottom row of Figure 7. For most of the sample periods, the estimated results from the two parametric methods are consistent with their whole sample estimates (Model 2). However, interestingly, in the second half of 2018, the parameter estimates switch from Model 2 to Model 1. This result is consistent with our argument that there is a non-negligible probability that the ML methods points to Model 1. This partially explains the diverse empirical findings that we have in the current literature. Although it appears to be a sample sensitive outcome, the fundamental cause of the problem is the bi-modality of the log-likelihood of the two ML methods as discussed earlier.

From Figure 7, there seems to be a small structural break in the model parameters in March 2020 at the onset of COVID-19, although the changes are not dramatic. To avoid its impact on the forecasting evaluation, we compute the MSFEs and MAFEs from January 5, 2018 to February 29, 2020. The MSFEs and MAFEs of the four estimation methods for forecasting horizons from one day to 50 days are presented in Figure 8. The left (right) panels are based on MSFE (MAFE). Evidently, when the forecasting horizon is short, there is no substantial differences among the four estmation methods. The gaps among the lines increase substantially as the forecasting horizon extends to 50. The gaps are more visible for SPY than XLB. Over the longer horizons, for SPY, the loss function of the Whittle method is consistently the smallest, followed by MPL. For XLB, the loss function of the two ML methods follow closely of each other across all horizons. The performs of LWE and LWE(diff) are similar to each other and not as good as those of the parametric methods.

Figure 8: Mean squared forecast error and mean absolute forecast error: SPY


(a) SPY: MSFE

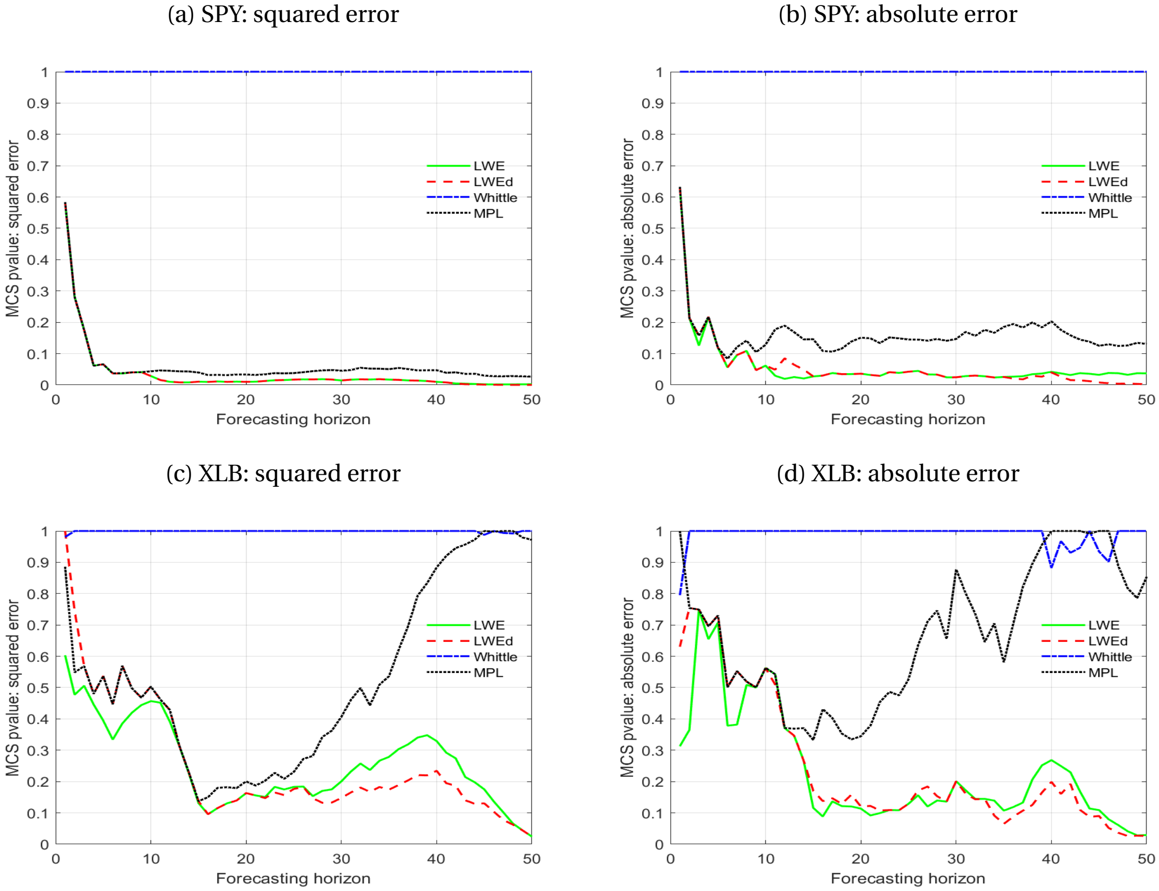
(b) SPY: MAFE

(c) XLB: MSFE

(d) XLB: MAFE

To investigate the statistical significance of those gaps, we conduct the MCS test. The MCS p-values are displayed in Figure 9. For SPY, the Whittle method always has the p-value of one, suggesting that it is the best model out of the four competing ones. With a significance level of 10%, we cannot reject all models being in the 'best model set' at the short forecasting horizon (up to 5 periods). Nevertheless, the Whittle method is the only survival model for horizons beyond five with the squared error loss function. With the absolute errors, at the 10% level, both MPL and Whittle survive the test and we do not have sufficient information to distinguish between these two methods. For XLB, at the 10%, all four methods survive from one period-ahead to a very long horizon forecast (45-period-ahead). The p-values of the Whittle method have almost consistently been the highest. For the remaining forecasting horizons, both MPL and Whittle have p-value close to one, while the p-values of the two semi-parametric methods are below 10%. The overall conclusion is that the Whittle method yields the most accurate forecasts, followed by MPL.

Figure 9: The MCS p-values



(a) SPY: squared error

(b) SPY: absolute error

(c) XLB: squared error

(d) XLB: absolute error

7 Conclusion

In this paper, we first examine the finite sample properties of four alternative methods in estimating the AR1FI(α, d) model, including the two parametric ML (MPL and Whittle) methods and two semi-parametric (LWE and LWE(diff)) methods. Special attention is paid to the part of the parameter space where d is close to -0.5 and α is close to unity (Model 2) and where d is close to 0.5 and α is close to zero (Model 1). These choices of parameter settings are motivated by the empirical findings in the RV literature. Via simulations, we find that all four methods have finite sample problems, although the problem associated with the ML methods is less severe. Specifically, LWE and LWE(diff) are significantly biased when the autoregressive coefficient α deviates far from zero and unity, respectively. Moreover, when the DGP is Model 2, LWE always points to Model 1. When the DGP is Model 1, LWE(diff) always points to Model 2. The two ML methods generally perform well. However, there exists a non-negligible probability that the ML methods mix up Model 1 and 2. This problem of the ML methods has never been discovered in the literature. The source of the problem is that the AR1FI$(0, d)$ model is observationally equivalent to the AR1FI$(1, d-1)$ model, leading to a weak identification problem. These simulation findings explain the contradicting empirical evidence documented in the literature.

We apply the four estimation methods to the log RVs of ten financial assets. The two ML methods and LWE(diff) always suggest Model 2, while LWE always suggests Model 1. Based on what we learn from the simulation studies, we conclude that the true DGP is more likely to be Model 2 than Model 1. Unfortunately, due to the aforementioned finite sample issue of the ML methods, we cannot draw definitive conclusions regarding the DGP. Despite the uncertainty, we find that the estimated model from the Whittle method can generate the most accurate out-of-sample forecasts for the log RV series, especially at long horizons.

References

An, S. and P. Bloomfield (1993). Cox and reid's modification in regression models with correlated errors. *Department of Statistics, North Carolina State University, Raleigh*.

Andersen, T. G. and T. Bollerslev (1997). Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. *The Journal of Finance 52*(3), 975–1005.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001). The distribution of realized stock return volatility. *Journal of Financial Economics 61*(1), 43–76.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of realized exchange rate volatility. *Journal of the American statistical association 96*(453), 42–55.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica 71*(2), 579–625.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and J. Wu (2005). A framework for exploring the macroeconomic determinants of systematic risk. *American Economic Review 95*(2), 398–404.

Baillie, R. T., F. Calonaci, D. Cho, and S. Rho (2019). Long memory, realized volatility and heterogeneous autoregressive models. *Journal of Time Series Analysis 40*(4), 609–628.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2009). Realized kernels in practice: Trades and quotes.

Barndorff-Nielsen, O. E. and N. Shephard (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of financial econometrics 2*(1), 1–37.

Beran, J. (1994). *Statistics for Long-Memory Processes*, Volume 61. CRC Press.

Bertelli, S. and M. Caporin (2002). A note on calculating autocovariances of long-memory processes. *Journal of Time Series Analysis 23*(5), 503–508.

Bloomfield, P. (1985). On series representations for linear predictors. *The Annals of Probability*, 226–233.

Bolko, A. E., K. Christensen, M. S. Pakkanen, and B. Veliyev (2021). Roughness in spot variance? a gmm approach for estimation of fractional log-normal stochastic volatility models using realized measures. *arXiv preprint arXiv:2010.04610*.

Brockwell, P. J. and R. A. Davis (1987). *Time Series: Theory and Methods.* Springer, New York.

Brockwell, P. J. and R. A. Davis (2009). *Time series: theory and methods.* Springer Science & Business Media.

Brownlees, C. T. and G. M. Gallo (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational statistics & data analysis 51*(4), 2232–2245.

Christoffersen, P., B. Feunou, K. Jacobs, and N. Meddahi (2014). The economic value of realized volatility: Using high-frequency returns for option valuation. *Journal of Financial and Quantitative Analysis*, 663–697.

Christoffersen, P. F. and F. X. Diebold (2000). How relevant is volatility forecasting for financial risk management? *Review of Economics and Statistics 82*(1), 12–22.

Chung, C.-F. (1994). A note on calculating the autocovariances of the fractionally integrated arma models. *Economics Letters 45*(3), 293–297.

Coursol, J. and D. Dacunha-Castelle (1982). Remarks on the approximation of the likelihood function of a stationary gaussian process. *Theory of Probability & Its Applications 27*(1), 162–167.

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological) 49*(1), 1–18.

Da, R. and D. Xiu (2021). When moving-average models meet high-frequency data: Uniform inference on volatility. *Econometrica* (17-27).

Dahlhaus, R. (1988). Small sample effects in time series analysis: a new asymptotic theory and a new estimate. *The Annals of Statistics*, 808–841.

Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *The Annals of Statistics*, 1749–1766.

Davydov, Y. A. (1970). The invariance principle for stationary processes. *Theory of Probability & Its Applications 15*(3), 487–498.

Diebold, F. X. (2003). The ET interview: Professor robert f. engle, january 2003. *Econometric Theory 19*(6), 1159–1193.

Duffie, D. and R. Kan (1996). A yield-factor model of interest rates. *Mathematical Finance 6*(4), 379–406.

Fleming, J., C. Kirby, and B. Ostdiek (2003). The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics 67*(3), 473–509.

Fox, R. and M. S. Taqqu (1986). Large-sample properties of parameter estimates for strongly dependent stationary gaussian time series. *The Annals of Statistics*, 517–532.

Fukasawa, M., T. Takabatake, and R. Westphal (2021). Consistent estimation for fractional stochastic volatility model under high-frequency asymptotics. *Mathematical Finance, forthcoming*.

Gatheral, J., T. Jaisson, and M. Rosenbaum (2018). Volatility is rough. *Quantitative Finance 18*(6), 933–949.

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

Giraitis, L. and D. Surgailis (1990). A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of whittle's estimate. *Probability Theory and Related Fields 86*(1), 87–104.

Granger, C. W. and R. Joyeux (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis 1*(1), 15–29.

Hannan, E. J. (1973). The asymptotic theory of linear time-series models. *Journal of Applied Probability*, 130–145.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Hauser, M. A. (1999). Maximum likelihood estimators for arma and arfima models: A monte carlo study. *Journal of Statistical Planning and Inference 80*(1-2), 229–255.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies 6*(2), 327–343.

Hosking, J. R. (1981). Fractional differencing. *Biometrika 68*(1), 165–76.

Hull, J. and A. White (1987). The pricing of options on assets with stochastic volatilities. *The Journal of Finance 42*(2), 281–300.

Hurvich, C. M. and W. W. Chen (2000). An efficient taper for potentially overdifferenced long-memory time series. *Journal of Time Series Analysis 21*(2), 155–180.

Hurvich, C. M. and B. K. Ray (1995). Estimation of the memory parameter for nonstationary or noninvertible fractionally integrated processes. *Journal of Time Series Analysis 16*(1), 17–41.

Jacod, J., Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic processes and their applications 119*(7), 2249–2276.

Jacod, J., Y. Li, and X. Zheng (2017). Statistical properties of microstructure noise. *Econometrica 85*(4), 1133–1174.

Künsch, H. (1987). Statistical aspects of self-similar processes. In *Proceedings of the First Congress of the Bernoulli Society, 1987.*

Lieberman, O. (2005). On plug-in estimation of long memory models. *Econometric Theory*, 431–454.

Liu, X., S. Shi, and J. Yu (2020). Persistent and rough volatility. *Available at SSRN 3724733*.

Magdalinos, T. (2012). Mildly explosive autoregression under weak and strong dependence. *Journal of Econometrics 169*(2), 179–187.

Meddahi, N. (2003). Arma representation of integrated and realized variances. *The Econometrics Journal 6*(2), 335–356.

Nadarajah, K., G. M. Martin, and D. Poskitt (2021). Optimal bias correction of the log-periodogram estimator of the fractional parameter: A jackknife approach. *Journal of Statistical Planning and Inference 211*, 41–79.

Nielsen, M. r. and P. H. Frederiksen (2005). Finite sample comparison of parametric, semiparametric, and wavelet estimators of fractional integration. *Econometric Reviews 24*(4), 405–443.

Phillips, P. C. and J. Yu (2009). A two-stage realized volatility approach to estimation of diffusion processes with discrete data. *Journal of Econometrics 150*(2), 139–150.

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

Phillips, P. C. B. and K. Shimotsu (2004). Local Whittle estimation in nonstationary and unit root cases. *The Annals of Statistics 32*(2), 656 – 692.

Robinson, P. M. (1986). On the errors-in-variables problem for time series. *Journal of Multivariate Analysis 19*(2), 240–250.

Robinson, P. M. (1995a). Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics 23*(5), 1630–1661.

Robinson, P. M. (1995b). Log-periodogram regression of time series with long range dependence. *The Annals of Statistics 23*(4), 1048–1072.

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

Shimotsu, K. and P. C. B. Phillips (2006). Local whittle estimation of fractional integration and some of its variants. *Journal of Econometrics 130*(2), 209–233.

Smith, J., N. Taylor, and S. Yadav (1997). Comparing the bias and misspecification in arfima models. *Journal of Time Series Analysis 18*(5), 507–527.

Sowell, F. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics 53*(1-3), 165–188.

Tanaka, K. (2013). Distributions of the maximum likelihood and minimum contrast estimators associated with the fractional ornstein–uhlenbeck process. *Statistical Inference for Stochastic Processes 16*(3), 173–192.

Tao, Y., P. C. Phillips, and J. Yu (2019). Random coefficient continuous systems: Testing for extreme sample path behavior. *Journal of Econometrics 209*(2), 208–237.

Tukey, J. W. (1967). An intoroduction to the calculation of numerical spectrum analysis. *Spectra Analysis of Time Series*, 25–46.

Varneskov, R. T. (2017). Estimating the quadratic variation spectrum of noisy asset prices using generalized flat-top realized kernels. *Econometric Theory 33*(6), 1457–1501.

Velasco, C. (1999). Gaussian semiparametric estimation of non-stationary time series. *Journal of Time Series Analysis 20*(1), 87–127.

Velasco, C. and P. M. Robinson (2000). Whittle pseudo-maximum likelihood estimation for nonstationary time series. *Journal of the American Statistical Association 95*(452), 1229–1243.

Wang, X., W. Xiao, and J. Yu (2021). Modeling and forecasting realized volatility with the fractional ornstein-uhlenbeck process. *Journal of Econometrics, forthcoming*.

Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv för matematik 2*(5), 423–434.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 434–449.

Yajima, Y. (1985). On estimation of long-memory time series models. *Australian Journal of Statistics 27*(3), 303–320.

Yu, J. (2021). Latent local-to-unity models. *SMU Economics and Statistics Working Paper Series, Paper No. 04-2021*.

Žurbenko, I. G. (1979). On the efficiency of estimates of a spectral density. *Scandinavian Journal of Statistics*, 49–56.

A Tapering

One pitfall of the periodogram $I\left(\lambda_j\right)$ is that there is leakage effect. In finite samples, when there are high peaks in the spectrum, the nonparametric estimator $I\left(\lambda_j\right)$ might significantly overestimate the spectrum at other frequencies and fail to discover spectrums with low peaks.

A.1 Whittle Estimator with Tapering

Dahlhaus (1988) proposes using tappering adapted from nonparametric spectral density estimation (Tukey, 1967) for the Whittle estimator. A tapered series is define as

$$y_t^T = h_t y_t,$$

where h_t is the data taper satisfying certain time series properties (Dahlhaus, 1988). The tapered periodogram is

$$I_T\left(\lambda_j\right) = \frac{1}{2\pi \sum_{t=0}^{T-1} h_t^2} \left| \sum_{t=0}^{T} h_t y_t \exp\left(-it\lambda_j\right) \right|^2 .$$

Replacing $I\left(\lambda_j\right)$ in the Whittle estimator (19) by $I_T\left(\lambda_j\right)$ yields the tapered Whittle estimator. Dahlhaus (1988) show that the tapered Whittle estimator is $\sqrt{T}$-consistent and asymptotically normal.

There are many tapers satisfying the conditions outlined in Dahlhaus (1988). One example is the Tukey-Hanning taper specified as

$$h_\rho\left(x\right) = \begin{cases} \frac{1}{2}\left[1-\cos\left(2\pi x/\rho\right)\right] & x \in [0, \rho/2) \\ 1 & x \in [\rho/2, 1/2] \\ h_\rho\left(1-x\right) & x \in (1/2, 1] \end{cases}$$

and $h_t = h_\rho\left(t/T\right)$. For practical implementation, one could set $\rho = T^{-\kappa/3}$ with $\kappa \in [0, 1/2)$. Here, we set $\kappa = 1/4$.

A.2 Local Whittle Estimator with Tapering

One popular tapering method in the local Whittle content is proposed by Velasco (1999). For each positive integer p, there is a Kolmogorov taper which is of order p in the sense of Velasco (1999). A taper with order p, if applied to the raw data, yields a tapered periodogram that is invariant to polynomial trends of order $p-1$, provided that the periodogram is evaluated on the grid λ_{ip} with $i = 1, 2, \ldots, \lfloor m/p \rfloor$. The objective function of the tapered LW estimator becomes

$$(\hat{C}_p, \hat{d}_p) = \arg\max_{C,d} \frac{p}{m} \sum_{i=1}^{\lfloor m/p \rfloor} \left[-\log C + 2d \log \lambda_{ip} - \frac{1}{C} \lambda_{ip}^{2d} I^T \left(\lambda_{ip}\right) \right], \tag{27}$$

where

$$\begin{aligned} \hat{d}_p &= \arg\max \left\{ -\log \hat{C}_p\left(d\right) + 2d \frac{p}{m} \sum_{i=1}^{\lfloor m/p \rfloor} \log \lambda_{ip} \right\}, \\ \hat{C}_p\left(d\right) &= \frac{p}{m} \sum_{i=1}^{\lfloor m/p \rfloor} \lambda_{ip}^{2d} I_T\left(\lambda_{ip}\right). \end{aligned}$$

The discrete sums include only frequencies λ_{ip} with $i = 1, 2, \ldots, \lfloor m/p \rfloor$.

The tapered LW estimator is asymptotic normal with a variance of $p\Phi/(4m)$, where

$$\Phi = \lim_{T\to\infty} \left(\sum_{t=1}^{T} h_t^2 \right)^{-2} \sum_{k=0,p,2p,\ldots}^{n-p} \left\{ \sum_{t=1}^{n} h_t^2 \cos\left(t\lambda_k\right) \right\}^2.$$

Suppose we employ the full cosine bell taper (Tukey, 1967)

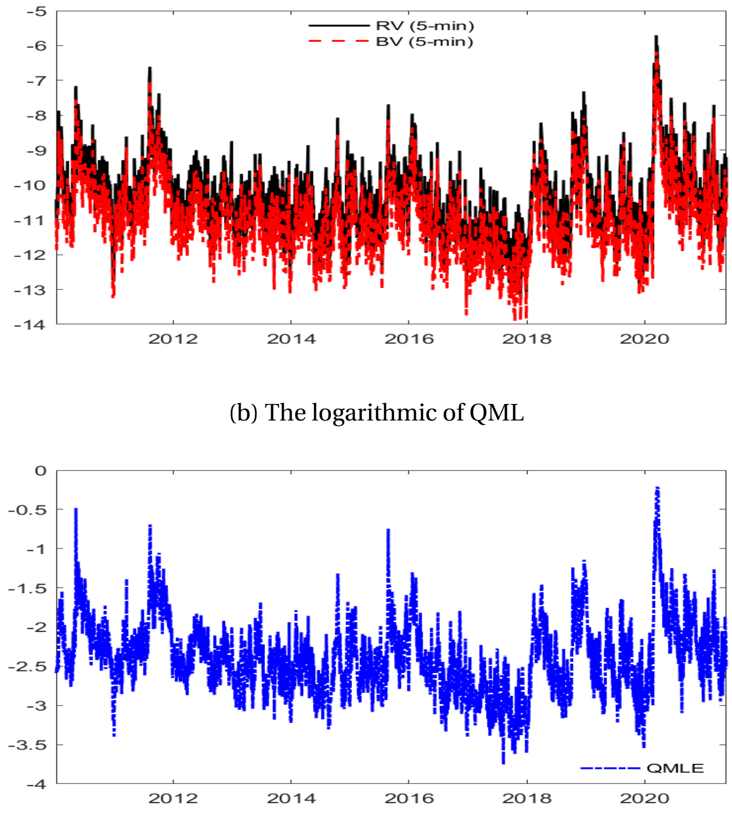
$$h_t = 0.5 \left[1 - \cos\left(\frac{2\pi t}{T} \right) \right]$$

and regard this taper as of order $p = 3$, the tapered LW estimator is asymptotic normal with variance $pm\Phi/4$ with $\Phi = 1$, when $\mu = 0$ and $d < 1.5$, . However, if we use all the Fourier frequencies from λ_2 to λ_m (i.e., $p = 1$), then $\Phi = 35/18$. In the simulation studies, we use the cosine bell taper with $p = 3$. While the tapered local Whittle methods are invariant to trends and asymptotically normal, they lead to inflated asymptotic variance of the estimator.

B Robustness Check: Alternative Estimators of Volatility

Taking the S&P 500 market index ETF as an example, we investigate the estimation robustness with respect to different measures of volatility. In addition to the QML estimator, we consider the popular realized volatility estimator as defined in (1) and the jump-robust volatility estimator – bipower variation of Barndorff-Nielsen and Shephard (2004). The data are downloaded from Refinitive Tick History at the one second frequency and sampled every five minutes. The 5-minute data are cleaned following the standard practice (Brownlees and Gallo, 2006; Barndorff-Nielsen et al., 2009). Figure 10 displays the estimated log volatilities over the sample period.

Figure 10: Volatility dynamics: the S&P500 market ETF


(a) The logarithmic of RV and BV

(b) The logarithmic of QML

Estimation results from the three data series are presented in Table 9. As for QML, the LWE of d for the logarithmic RV and BV are higher than 0.5. The two parametric methods lead to similar estimation results. For both data series, the estimated d is negative and around -0.475, which is smaller than that from log RV (QML). For all data series, the estimated autoregressive coefficients are all very close to unity, with the one from log RV (QML) being the smallest. The LWE(diff) method leads to similar estimation results. As expected, the estimated fractional parameters are slightly higher than those from the parametric methods. In summary, the parametric methods point to Model 2 for all three volatility series.

Table 9: Model estimation results: S&P 500 ETF

Estimator	MPL		Whittle		LWE		LWE(diff)
	$\hat{d}$	$\hat{\alpha}$	$\hat{d}$	$\hat{\alpha}$	$\hat{d}$	$\hat{\alpha}$	$\hat{d}$
log RV	-0.480	0.998	-0.480	0.997	0.542	-0.058	-0.445
log BV	-0.472	0.997	-0.472	0.997	0.539	-0.043	-0.438
log RV (QML)	-0.382	0.995	-0.382	0.995	0.602	-0.006	-0.356