Journal of Econometrics 150 (2009) 139–150

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A two-stage realized volatility approach to estimation of diffusion processes with discrete data☆

Peter C.B. Phillips [a,b,c,d], Jun Yu [e,*]

[a] Cowles Foundation, Yale University, United States
[b] Department of Economics, University of Auckland, New Zealand
[c] Department of Economics, University of York, United Kingdom
[d] School of Economics, Singapore Management University, Singapore
[e] School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903, Singapore

ARTICLE INFO

Article history:
Available online 25 December 2008

JEL classification:
C13
C22
E43
G13

Keywords:
Maximum likelihood
Girsanov theorem
Discrete sampling
Continuous record
Realized volatility

ABSTRACT

This paper motivates and introduces a two-stage method of estimating diffusion processes based on discretely sampled observations. In the first stage we make use of the feasible central limit theory for realized volatility, as developed in [Jacod, J., 1994. Limit of random measures associated with the increments of a Brownian semiartingal. Working paper, Laboratoire de Probabilities, Universite Pierre et Marie Curie, Paris] and [Barndorff-Nielsen, O., Shephard, N., 2002. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society. Series B, 64, 253–280], to provide a regression model for estimating the parameters in the diffusion function. In the second stage, the in-fill likelihood function is derived by means of the Girsanov theorem and then used to estimate the parameters in the drift function. Consistency and asymptotic distribution theory for these estimates are established in various contexts. The finite sample performance of the proposed method is compared with that of the approximate maximum likelihood method of [Aït-Sahalia, Y., 2002. Maximum likelihood estimation of discretely sampled diffusion: A closed-form approximation approach. Econometrica. 70, 223–262].

1. Introduction

For many years, continuous time models have enjoyed a great deal of success in finance (Merton, 1990) as well as wide applications in economics (e.g., Dixit (1993)). Correspondingly, there has been growing interest in estimating continuous systems using econometric methods with discrete data.

Many models used in finance for modelling asset prices can be written in terms of a diffusion process as

$$dX_t = \mu(X_t; \theta_1)dt + \sigma(X_t; \theta_2)dB_t, \tag{1}$$

where B_t is a standard Brownian motion, $\sigma(X_t; \theta_2)$ is a known diffusion function, $\mu(X_t; \theta_1)$ is a known drift function, and $\theta = (\theta_1, \theta_2)'$ is a vector of $k_1 + k_2$ unknown parameters. Note that we isolate the vector of parameters θ_2 in the diffusion function from θ_1 for reasons which will be clear below. The attractions of the Ito calculus make it easy to work with processes generated by diffusions like (1) and as a result these processes have been used widely in finance to model asset prices, including stock prices, interest rates, and exchange rates.

From an econometric standpoint, the estimation problem is to estimate θ from observed data, which are typically recorded discretely at $(\Delta, 2\Delta, \ldots, n_\Delta\Delta(\equiv T))$ over a certain time interval $[0, T]$, where Δ is the sampling interval and T is the time span of the data. For example, if X_t is recorded as the annualized interest rate and observed monthly (weekly or daily), we have $\Delta = 1/12$ (1/52 or 1/250). Typically, T can be as large as 50 for US Treasury Bills, but is generally much smaller for data from swap

☆ We thank two anonymous referees for their constructive comments. We also thank Yacine Aït-Sahalia, Yongmiao Hong, Chung-ming Kuan, Andy Lo, and Neil Shephard, and seminar participants at the First Symposium on Econometric Theory and Applications in Taiwan, the 2006 North American Winter Meeting of Econometric Society, and the First Finance Summer Camp at Singapore Management University for helpful discussions. Phillips gratefully acknowledges visiting support from the School of Economics at Singapore Management University, and support from a Kelly Fellowship at the University of Auckland Business School and from the NSF under Grant No. SES 04-142254. Yu gratefully acknowledges financial support from the Ministry of Education AcRF Tier 2 fund under Grant No. T206B4301-RS.

* Corresponding author.
E-mail addresses: peter.phillips@yale.edu (P.C.B. Phillips), yujun@smu.edu.sg (J. Yu).

0304-4076/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.12.006

markets. Also, due to time-of-day effects and possibly other market microstructure frictions, it is commonly believed that intra-day data do not completely follow diffusion models such as (1). As a result, daily and lower frequencies are most frequently used to estimate continuous time models. However, Barndorff-Nielsen and Shephard (2002) and Bollerslev and Zhou (2002) recently showed how to use information from intra-day data to estimate continuous time stochastic volatility models.

A large class of estimation methods is based on the likelihood function derived from the transition probability density of discrete sampling and then resorts to long span asymptotic theory (i.e. $T \to \infty$). Except for a few cases, the transition probability density does not have a closed form expression and hence the exact maximum likelihood (ML) method based on the likelihood function for the discretely sampled data is not directly available. In the financial econometrics literature, interest in obtaining estimators which approximate or approach ML estimators has been growing, in view of the natural attractiveness of maximum likelihood and its asymptotic properties. Several alternative methods of this type have been developed in recent years. See Phillips and Yu (in press-a) for a survey of various alternative methods and the discussion of their advantages and drawbacks.

The main purpose of the present paper is to propose an alternative method of estimating diffusion processes of the form given by model (1) from discrete observations and to establish asymptotic properties by resorting to both the long span (i.e. $T \to \infty$) and in-fill asymptotics (i.e. $\Delta \to 0$). The estimation procedure involves two steps. In the first step, we propose to use a quadratic variation type estimator of θ_2. In the second step, an approximate in-fill likelihood function is maximized to obtain a ML estimator of θ_1. This method has several advantages over the existing method. First, it is not dependent on finding an appropriate auxiliary model. Second, it does not require simulations or polynomial expansions and hence is straightforward to implement. Third, it decomposes the optimization problem into two smaller scale optimization problems, making the approach computationally more attractive. Finally, experience with the procedure both in simulations and empirical applications indicates that the method works well in finite samples.

The paper is organized as follows: Section 2 reviews the literature on the ML estimation of diffusion processes and motivates the approach. Section 3 introduces the new method and Section 4 derives the asymptotic properties of the estimates. Section 5 presents some Monte Carlo evidence. Section 6 discusses the case of microstructure noises and Section 6 concludes. Proofs are provided in the Appendix.

2. Literature review and motivation

2.1. Literature review

2.1.1. Transition probability density based approaches

As explained above, a large class of estimation methods is based on the likelihood function derived from the transition probability density of the discretely sampled data. Suppose $p(X_{i\Delta}|X_{(i-1)\Delta}, \theta)$ is the transition probability density. The Markov property of model (1) implies the following log-likelihood function for the discrete sample

$$\ell_{TD}(\theta) = \sum_{i=2}^{n_\Delta} \log(p(X_{i\Delta}|X_{(i-1)\Delta}, \theta)). \tag{2}$$

Under regularity conditions, the resulting estimator is consistent, asymptotically normally distributed and asymptotically efficient (Billingsley, 1961). Unfortunately, except for a few cases, the transition density does not have a closed form expression and hence the exact ML method based on the likelihood function of the discrete sample is not a practical procedure. In the financial econometrics literature, interest in finding estimators that approach ML estimators in some quantifiable sense has been growing and many alternative methods have been developed in recent years. For example, Lo (1988) suggested calculating the transition probability density by solving a partial differential equation numerically. Nowman (1997) suggested an approach which assumes that the conditional volatility remains unchanged over the unit intervals so that he can approximate the transition density using a Gaussian method. Yu and Phillips (2001) used the stoping time technique to develop an exact Gaussian method. Pedersen (1995) and Brandt and Santa-Clara (2002) advocated an approach which calculates the transition probability density using simulation with some auxiliary points between each pair of consecutive observations introduced. This method is also closely related to the Bayesian MCMC method proposed by Elerian et al. (2001) and Eraker (2001). A drawback of these simulation-based approaches is that the corresponding computational cost will inevitably be high.

As an important alternative to these numerical and simulated ML methods, Aït-Sahalia (2002) proposed to approximate the transition probability density of diffusions using analytical expansions via Hermite polynomials. Before obtaining the closed-form expansions, a Lamperti transform is performed on the continuous time model so that the diffusion function becomes a constant. After that one then obtains a Hermite polynomial expansion of the transition density of the transformed variable around the normal distribution. Aït-Sahalia (1999) implemented the approximate ML method and documents its good performance. As it is typically tedious and error prone to derive the Hermite expansion by hand, Aït-Sahalia (2002) suggested using symbolic softwares, such as MATHEMATICA.

Apart from these likelihood-based approaches, numerous alternative methods are available. We simply refer readers to the book by Prakasa Rao (1999a) for a review of many alternative approaches.

2.1.2. Approaches based on realized volatility and in-fill likelihood

When the transition probability density does not have a closed form expression but X_t is observed continuously over $[0, T]$, an alternative method can be used to estimate diffusion models. We now introduce and motivate the approach.

When the diffusion term is known (i.e. $\sigma(X_t; \theta_2) = \sigma(X_t)$) and so does not depend on any unknown parameters, one can construct the exact continuous record log-likelihood via the Girsanov theorem (e.g., (Liptser and Shiryaev, 2000) as follows.

$$\ell_{IF}(\theta_1) = \int_0^T \frac{\mu(X_t; \theta_1)}{\sigma^2(X_t)} \mathrm{d}X_t - \frac{1}{2} \int_0^T \frac{\mu^2(X_t; \theta_1)}{\sigma^2(X_t)} \mathrm{d}t.$$

Lánska (1979) established the consistency and asymptotic normality of the continuous record ML estimator of θ_1 when $T \to \infty$ under a certain set of regularity conditions.

The assumptions of a known diffusion function and the availability of a continuous time record are not realistic in financial and other applications. Motivated by the fact that the drift and diffusion functions are of different orders (Bandi and Phillips, 2003, 2007), we argue that there can be advantages to estimating the diffusion parameters separately from the drift parameters. For example, when $\sigma(X_t; \theta_2) = \theta_2$, i.e., the diffusion function is an unknown constant, a two-stage approach can be used to estimate the model. First, θ_2 can be estimated directly by the realized volatility function, i.e.,

$$\hat{\theta}_2 = \sqrt{\frac{[X_\Delta]_T}{T}}, \tag{3}$$

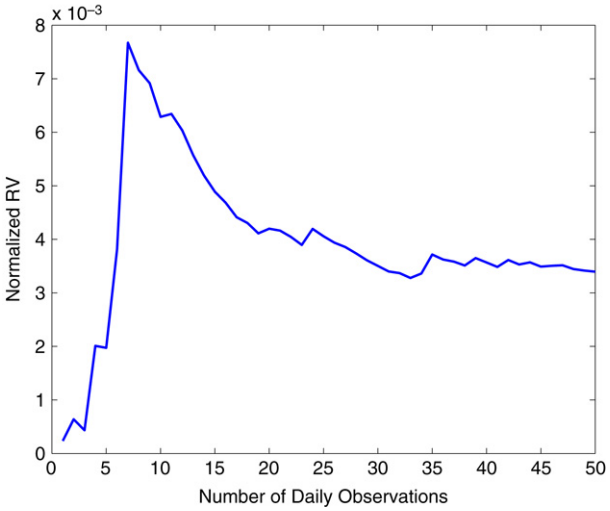

Fig. 1. Standardized realized volatility against the number of daily observations used to calculate the realized volatility. The daily data are simulated from the Vasicek model $dX_t = 0.6(0.09 - X_t)dt + 0.06dB_t$.

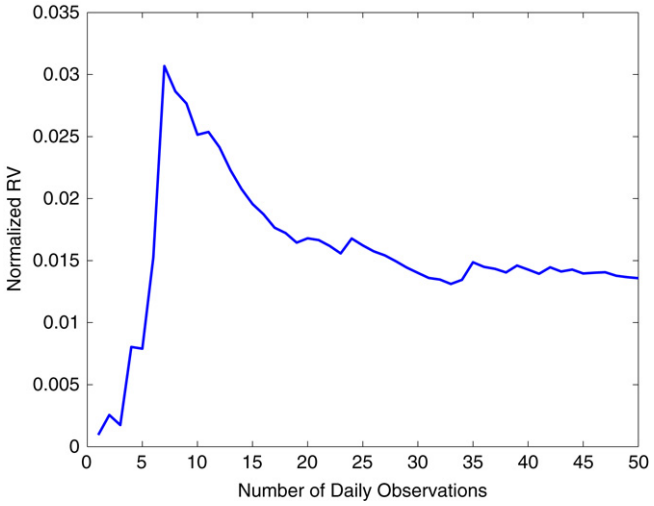

Fig. 2. Standardized realized volatility against the number of daily observations used to calculate the realized volatility. The daily data are simulated from the Vasicek model $dX_t = 0.6(0.09 - X_t)dt + 0.12dB_t$.

where $[X_\Delta]_T = \sum_{i=2}^{n_\Delta}(X_{i\Delta} - X_{(i-1)\Delta})^2$. This is because model (1) implies that

$$(dX_t)^2 = \theta_2^2 dt, \quad \forall t,$$

and hence

$$[X]_T = \int_0^T (dX_t)^2 = \int_0^T \theta_2^2 dt = T\theta_2^2,$$

where $[X]_T$ is the quadratic variation of X, which can be consistently estimated by $[X_\Delta]_T$ as $\Delta \to 0$. As a result, $\hat{\theta}_2$ should be a very good estimate of θ_2 when Δ is small, which is typically the case for interest rate data. Indeed, in the special case where the drift term is zero and the diffusion term is an unknown constant, the exact discrete model (Phillips, 1972) for the data is $X_{i\Delta} - X_{(i-1)\Delta} = \theta_2 \left(B_{i\Delta} - B_{(i-1)\Delta}\right)$ and so the maximum likelihood estimator is trivially $\hat{\theta}_2$ (see also Aït-Sahalia et al. (2005)). Although this correspondence clearly does not apply to more general specifications, it seems likely that when Δ is small the two estimators will be close to each other. Second, the following logarithmic continuous record likelihood function of model (1),

$$\ell_{IF}(\theta_1) = \int_0^T \frac{\mu(X_t;\theta_1)}{\sigma^2(X_t;\hat{\theta}_2)} dX_t - \frac{1}{2}\int_0^T \frac{\mu^2(X_t;\theta_1)}{\sigma^2(X_t;\hat{\theta}_2)} dt,$$

may be approximated by the in-fill likelihood function

$$\ell_{AIF}(\theta_1) = \sum_{i=2}^{n_\Delta} \frac{\mu(X_{(i-1)\Delta};\theta_1)}{\hat{\theta}_2^2}(X_{i\Delta} - X_{(i-1)\Delta}) - \frac{\Delta}{2}\sum_{i=2}^{n_\Delta} \frac{\mu^2(X_{(i-1)\Delta};\theta_1)}{\hat{\theta}_2^2}, \tag{4}$$

which is in turn maximized with respect to θ_1. Because θ_1 is estimated by ML and $\widehat{\theta}_2$ is close to an MLE, the two-stage procedure may be interpreted as a form of profile ML estimation. This two-stage approach is closely related to the method proposed by Florens-Zmirou (1989), where a contrast function instead of the logarithmic in-fill likelihood function was used in the second step.

To better appreciate the quality of approximation of $[X_\Delta]_T$ to $[X]_T$, we simulate daily data from the Vasicek model

$$dX_t = 0.6(0.09 - X_t)dt + 0.06dB_t, \tag{5}$$

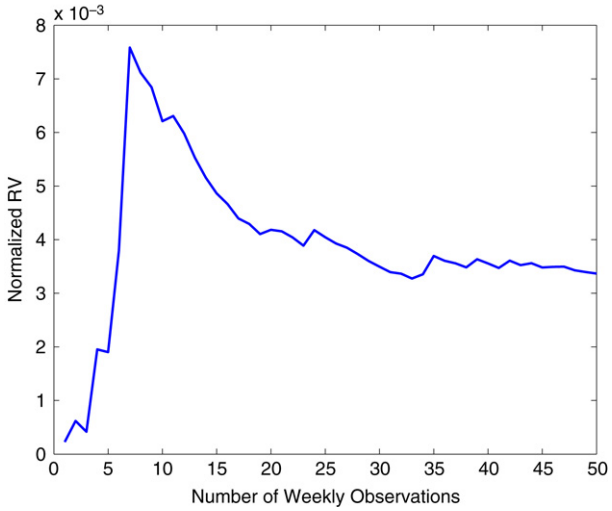

Fig. 3. Standardized realized volatility against the number of weekly observations used to calculate the realized volatility. The weekly data are simulated from the Vasicek model $dX_t = 0.6(0.09 - X_t)dt + 0.06dB_t$.

and plot the standardized realized volatility, $[X_\Delta]_T/T$, against the number of daily observations used to calculate $[X_\Delta]_T$. Quadratic variation theory implies that as $\Delta \to 0$, $[X_\Delta]_T/T \xrightarrow{a.s.} \sigma^2$. It is clear from Fig. 1 that although $[X_\Delta]_T/T$ is quite erratic initially but quickly settles down around σ^2. It seems that with only 30–35 observations one can get a good estimate of σ^2. We then increase the volatility rate from 0.06 to 0.12 and decrease the sampling frequency from daily to weekly. Results are plotted in Figs. 2 and 3, respectively. It clear that the conclusion about the rapid convergence of $[X_\Delta]_T/T$ is quite robust to these changes.

When the diffusion term is only known up to a scalar factor, that is when

$$dX_t = \mu(X_t;\theta_1)dt + \theta_2 f(X_t)dB_t, \tag{6}$$

the above two-stage method is easily modified. First, θ_2^2 can be estimated by

$$\hat{\theta}_2^2 = \frac{[X_\Delta]_T}{\Delta\sum_{i=2}^{n_\Delta} f^2(X_{(i-1)\Delta})}. \tag{7}$$

Table 1
Simulation results under the CIR model, $dX_t = 0.3(0.09 - X_t)dt + 0.06\sqrt{X_t}dB_t$, based on 1000 replications. Mean and SD stand for the average and standard deviation across 1000 replications, respectively.

Δ	T	Method	$\kappa = 0.3$		$\mu = 0.09$		$\sigma = 0.06$	
			Mean	SD	Mean	SD $\times 100$	Mean	SD $\times 100$
1/12	20	MLE	.5417	.2832	.0898	1.3848	.0603	.2841
		Yoshida	.5265	.2832	.0898	1.3785	.0598	.2779
		Proposed	.5265	.2663	.0898	1.3785	.0597	.2793
1/12	15	MLE	.6350	.3610	.0903	1.8937	.0604	.3232
		Yoshida	.6133	.3355	.0904	1.9611	.0597	.3150
		Proposed	.6133	.3355	.0904	1.9611	.0596	.3198
1/52	20	MLE	.5075	.2582	.0906	1.3467	.0601	.1332
		Yoshida	.5045	.2552	.0906	1.3470	.0600	.1326
		Proposed	.5045	.2552	.0906	1.3470	.0600	.1347
1/52	10	MLE	.7154	.4390	.0925	2.4234	.0601	.2035
		Yoshida	.7069	.4306	.0924	2.3301	.0600	.2020
		Proposed	.7069	.4306	.0924	2.3301	.0600	.2024
1/250	20	MLE	.5268	.2725	.0898	1.3176	.0600	.0617
		Yoshida	.5260	.2718	.0898	1.3179	.0600	.0617
		Proposed	.5260	.2718	.0898	1.3179	.0600	.0634
1/250	10	MLE	.7533	.4737	.0904	1.9306	.0601	.0874
		Yoshida	.7519	.4714	.0903	1.9283	.0600	.0877
		Proposed	.7519	.4714	.0903	1.9283	.0600	.0891

Second, the following approximate logarithmic in-fill likelihood function can then be maximized with respect to θ_1 (denoting the resulting estimator by $\hat{\theta}_1$)

$$\ell_{AIF}(\theta_1) = \sum_{i=2}^{n_\Delta} \frac{\mu(X_{(i-1)\Delta}; \theta_1)}{\hat{\theta}_2^2 f^2(X_{(i-1)\Delta})}(X_{i\Delta} - X_{(i-1)\Delta}) - \frac{\Delta}{2}\sum_{i=2}^{n_\Delta} \frac{\mu^2(X_{(i-1)\Delta}; \theta_1)}{\hat{\theta}_2^2 f^2(X_{(i-1)\Delta})}. \quad (8)$$

This method is applicable to many popular interest rate models, including those proposed by Vasicek (1977), Cox et al. (1985) (CIR hereafter), and Ahn and Gao (1999). It is also closely related to the method proposed by Yoshida (1992). In particular, instead of using the estimator in (7), Yoshida (1992) used the following estimator for θ_2^2:

$$\tilde{\theta}_2^2 = \frac{1}{T}\sum_{i=2}^{n_\Delta} \frac{(X_{i\Delta} - X_{(i-1)\Delta})^2}{f^2(X_{(i-1)\Delta})}. \quad (9)$$

Also,Yoshida (1992) suggested using an iterative procedure to construct a better estimate of θ_2^2 (denoted by $\tilde{\theta}_2^2$). Under the conditions of $\Delta \to 0$, $T \to \infty$, and $\Delta^2 T \to 0$, Yoshida (1992) derived limiting normal distributions for $\sqrt{n_\Delta}(\tilde{\theta}_2^2 - \theta_2^2)$ and $\sqrt{T}(\hat{\theta}_1 - \theta_1)$. Since $\sqrt{n_\Delta}/\sqrt{T} = \sqrt{1/\Delta} \to \infty$, the diffusion parameter enjoys a faster rate of convergence. Unfortunately, requiring the diffusion function to be either a constant or known up to a scalar function limits applicability and the procedures proposed by Florens-Zmirou (1989) and Yoshida (1992) cannot be implemented with more general diffusion processes.

The restriction on the diffusion term regarding parameter dependence was somewhat "relaxed" in Hutton and Nelson (1986) who based estimation on the following first order condition of the logarithmic quasi-likelihood function:

$$\int_0^T \frac{\partial\mu(X_t;\theta)/\partial\theta}{\sigma^2(X_t;\theta)}dX_t - \frac{1}{2}\int_0^T \frac{\partial\mu^2(X_t;\theta)/\partial\theta}{\sigma^2(X_t;\theta)}dt = 0.$$

Although their model seems to allow for a more flexible diffusion function, it requires that the drift term share the same set of parameters as the diffusion term. This assumption is too restrictive for practical applications. Moreover, although this one-stage estimation approach is easy to implement, the estimation is mainly based on the drift function and hence leads to inferior finite sample properties, as we will show below in the context of a simple example.

2.2. Motivation

Our two-stage method is in line with methods proposed by Florens-Zmirou (1989) andYoshida (1992). That is, in the first stage, we estimate the parameters in the diffusion functions based on the realized volatility, a quantity which consistently estimates the quadratic variation under very mild conditions. In the second step, by assuming the diffusion function to be known, we derive and approximate the logarithmic in-fill likelihood function. To motivate the two-step approach, we consider two simple examples.

2.2.1. Example 1

In the first example, we consider estimating the following CIR model

$$dX_t = \kappa(\mu - X_t)dt + \sigma\sqrt{X_t}dB_t, \quad (10)$$

using the exact ML method based on the transition probability density and the two-stage method discussed in Section 2.1.

The natural estimator of σ based on realized volatility is

$$\hat{\sigma} = \sqrt{\frac{[X_\Delta]_T}{\Delta\sum_{i=1}^{n_\Delta} X_{(i-1)\Delta}}}. \quad (11)$$

Moreover, since $\theta_1 = (\kappa, \mu)'$, the logarithmic in-fill likelihood is,

$$\sum_{i=1}^{n} \frac{\kappa(\mu - X_{(i-1)\Delta})(X_{i\Delta} - X_{(i-1)\Delta})}{\hat{\sigma}^2 X_{(i-1)\Delta}} - \frac{\Delta}{2}\sum_{i=1}^{n} \frac{\kappa^2(\mu - X_{(i-1)\Delta})^2}{\hat{\sigma}^2 X_{(i-1)\Delta}}. \quad (12)$$

CIR (1985) showed that the distribution of $X(t+\Delta)$ conditional on $X(t)$ is non-central chi-squared, $\chi^2[2cX(t), 2q + 2, 2\lambda(t)]$, where $c = 2\kappa/(\sigma^2(1 - e^{-\kappa\Delta}))$, $\lambda(t) = cr(t)e^{-\kappa\Delta}$, $q = 2\kappa\mu/\sigma^2 - 1$, and the second and third arguments are the degrees of freedom and non-centrality parameters, respectively. This transition probability density is used to calculate the likelihood function and to obtain the exact ML estimates.

Table 1 reports some results obtained from a Monte Carlo study where we compare three estimation methods: exact ML, Yoshida's method which estimates σ by (9), and the proposed method which estimates σ by (11). We vary both the sampling frequencies and

Table 2
Simulation results under $dX_t = 0.1dt + 0.1dB_t$ based on 1000 replications. Mean and variance are calculated across 1000 replications, respectively.

Δ	T	True value of $\alpha = 0.1$					
		RV		ML		QML	
		Mean	Variance $\times 100$	Mean	Variance $\times 100$	Mean	Variance $\times 100$
1/12	20	.1013	.0224	.1054	.0266	.0954	.469
1/52	20	.1003	.00488	.1013	.00514	.1005	.5121
1/250	20	.1000	.0011	.1002	.0011	.0992	.518

time spans. Note that the parameters and the sampling frequencies are all set to empirically reasonable values. In all cases, the two two-stage methods are almost identical. This is not surprising as the two methods differ only in the first stage. Moreover, the two-stage methods perform comparably with the ML method. Even in the case where very coarsely sampled data ($\Delta = 1/12$) are available, the two-stage method works quite well. In most cases the two-stage methods perform slightly better than the ML method. In light of Phillips and Yu (2005), the observed bias in the estimates of κ are the result of the near unit root problem. The observation that the two-stage method is not dominated by ML is quite remarkable, as the data generating process is based on the transition probability density on which ML itself is based. An interesting side result to emerge from this simulation is that the two-stage method is able to reduce the finite sample bias and variance in κ in all cases, even though the reductions are small.

2.2.2. Example 2

The model in the second example is taken from Hutton and Nelson (1986)

$$dX_t = \alpha dt + \alpha dB_t. \tag{13}$$

Although this model is generally not well suited to interest rate data, the feature that the drift and diffusion functions share the same parameter provides a nice framework to investigate the relative performance of the estimation method based on the diffusion only, against that based on the drift only and that based on the drift and diffusion jointly.

The first method is based on the realized volatility and hence only uses the diffusion term to estimate the model. It is easy to show that

$$\hat{\alpha}_1 = \sqrt{\frac{[X_\Delta]_T}{T}}.$$

The second method is based on the transition probability density given by

$$X_{i\Delta}|X_{(i-1)\Delta} \sim N(X_{(i-1)\Delta} + \alpha\Delta, \alpha^2\Delta).$$

Clearly this method uses information both in the drift and diffusion functions. Denote the resulting estimate by $\hat{\alpha}_2$. Note that this estimator is equivalent to the MLE of the discretized model via the Euler approximation.

The third method was proposed by Hutton and Nelson (1986). It uses mainly information in the drift function and is based on maximization of the following logarithmic quasi-likelihood function

$$\int_0^T \alpha^{-2} dX_t - \int_0^T \alpha^{-1} dt.$$

As a result, the estimate has the following analytical expression:

$$\hat{\alpha}_3 = \frac{X_T}{T}.$$

Table 2 reports results obtained from a Monte Carlo study where we compare the three estimation methods with different sampling frequencies. In all cases, the two-stage method and ML perform much better than QML; and, most remarkably, the two-stage method performs better than ML (and hence the Euler method). Just as in example 1, the fact that the simple two-stage method outperforms ML in finite samples is surprising. Moreover, the better performance of the first and second methods clearly reflects the order difference in the drift and diffusion functions.

3. A two-stage method

The estimation procedure discussed in Section 2.1 is not directly applicable to general diffusions such as model 1, as it requires either a constant diffusion function or separability of the scalar parameter from the reminder of the diffusion function. As a result, we have to provide a more general two-step procedure to estimate a diffusion process in the form of model (1)). In particular, in the first step we propose to estimate the parameters in the diffusion function by using the feasible central limit theorem for realized volatility derived by Jacod (1994) and popularized by Barndorff-Nielsen and Shephard (2002).

Assume that X_t is observed at a gird of discrete times

$$t = \Delta, 2\Delta, \ldots, M_\Delta\Delta \left(= \frac{T}{K}\right), (M_\Delta + 1)\Delta, \ldots,$$

$$2M_\Delta\Delta \left(= \frac{2T}{K}\right), \ldots, n_\Delta\Delta (= T),$$

where $n_\Delta = KM_\Delta$ with K being a fixed and positive integer, T is the time span of the data, Δ is the sampling frequency, and $M_\Delta = O(n_\Delta)$. This particular construction allows for the non-overlapping K sub-samples

$$((k-1)M_\Delta + 1)\Delta, \ldots, kM_\Delta\Delta, \quad \text{where } k = 1, \ldots, K,$$

so that each sub-sample has M_Δ observations over the interval $((k-1)\frac{T}{K}, k\frac{T}{K}]$. For example, if ten years of weekly observed interest rates are available and we split the data into ten blocks, then $T = 10$, $\Delta = 1/52$, $M_\Delta = 52$, $K = 10$. The total number of observations is 520 and the number of observations contained in each block is 52. The first limit in Box I follows by virtue of the definition of quadratic variation, while the second limit in Box I is the central limit theorem (CLT) which is due to Jacod (1994) and Barndorff-Nielsen and Shephard (2002), and the third limit in Box I involves a finite sample correction on the asymptotic theory (Barndorff-Nielsen and Shephard, 2005). It is shown in the latter reference that the third limit in Box I has a better finite sample performance than that the second limit in Box I.

Although in this paper we only use the realized volatility to estimate the quadratic variation, other realized power variations, such as realized absolute variation, can be used in the same way. For the theoretical development of general realized power variations, we refer readers to the articles by Barndorff-Nielsen and Shephard (2003) and Barndorff-Nielsen et al. (2006).

Based on the CLT i.e., the second limit in Box I, θ_2 can be estimated in the first stage by running a (nonlinear) least squares regression of the standardized realized volatility

$$\frac{\sum_{i=2}^{M_\Delta} (X_{(k-1)M_\Delta + i\Delta} - X_{(k-1)M_\Delta + (i-1)\Delta})^2}{r_k} \tag{14}$$

As $\Delta \to 0$, $n_\Delta = \frac{T}{\Delta} \to \infty$ and $M_\Delta \to \infty$, so that

$$\sum_{i=2}^{M_\Delta} (X_{(k-1)M_\Delta+i\Delta} - X_{(k-1)M_\Delta+(i-1)\Delta})^2 \xrightarrow{p} [X]_{k\frac{T}{K}} - [X]_{(k-1)\frac{T}{K}},$$

and

$$\frac{\sum_{i=2}^{M_\Delta} (X_{(k-1)M_\Delta+i\Delta} - X_{(k-1)M_\Delta+(i-1)\Delta})^2 - ([X]_{k\frac{T}{K}} - [X]_{(k-1)\frac{T}{K}})}{r_k} \xrightarrow{d} N(0,1),$$

$$\frac{\log\left(\sum_{i=2}^{M_\Delta} (X_{(k-1)M_\Delta+i\Delta} - X_{(k-1)M_\Delta+(i-1)\Delta})^2 - \log([X]_{k\frac{T}{K}} - [X]_{(k-1)\frac{T}{K}}) + \frac{1}{2}s_k^2\right)}{s_k} \xrightarrow{d} N(0,1),$$

where

$$r_k = \sqrt{\frac{2}{3}\sum_{i=2}^{M_\Delta} (X_{(k-1)M_\Delta+i\Delta} - X_{(k-1)M_\Delta+(i-1)\Delta})^4}$$

and

$$s_k = \max\left\{\sqrt{\frac{r_k^2}{(\sum_{i=2}^{M_\Delta} (X_{(k-1)M_\Delta+i\Delta} - X_{(k-1)M_\Delta+(i-1)\Delta})^2)^2}}, \sqrt{\frac{2}{M_\Delta}}\right\}$$

for $k = 1, \ldots, K$.

Box I.

$$\hat{\theta}_2 = \arg\min_{\theta_2} Q_\Delta(\theta_2),$$

where

$$Q_\Delta(\theta_2) = \Delta \sum_{k=1}^{K} \left[\frac{\sum_{i=2}^{M_\Delta} \left\{(X_{(k-1)M_\Delta+i\Delta} - X_{(k-1)M_\Delta+(i-1)\Delta})^2 - \sigma^2\left(X_{(k-1)M_\Delta+(i-1)\Delta}; \theta_2\right)\Delta\right\}}{r_k}\right]^2$$

Box II.

on the standardized diffusion function

$$\frac{([X]_{k\frac{T}{K}} - [X]_{(k-1)\frac{T}{K}})}{r_k} = \frac{\left(\int_{(k-1)\frac{T}{K}}^{k\frac{T}{K}} \sigma^2(X_t; \theta_2)\,\mathrm{d}t\right)}{r_k} \tag{15}$$

$$\simeq \frac{\sum_{i=2}^{M_\Delta} \sigma^2\left(X_{(k-1)M_\Delta+(i-1)\Delta}; \theta_2\right)\Delta}{r_k} \tag{16}$$

for $k = 1, \ldots, K$. Denote the resulting estimator of θ_2 by $\hat{\theta}_2$. In fact, we can write $\hat{\theta}_2$ as the extremum estimator as in Box II. A similar regression in standardized log levels of realized volatility can be run using the third limit in Box I.

This approach provides a more general estimation procedure than those designed to estimate models with a constant diffusion or a scalar parameter in the diffusion function. Hence this approach substantially generalizes the method of Florens-Zmirou (1989) and the method of Yoshida (1992) to a much wider class of diffusion processes. Indeed, when $K = 1$, the least squares regression above is equivalent to minimizing the squared difference between the terms given by Eqs. (14) and (15), which yields exactly the expression of the estimator (7) when the diffusion term is known up to the scalar factor.

In the second stage, the approximate log-likelihood function is maximized with respect to θ_1 (denoting the resulting estimator by $\hat{\theta}_1$)

$$\ell_{AIF}(\theta_1) = \sum_{i=2}^{n_\Delta} \frac{\mu(X_{(i-1)\Delta}; \theta_1)}{\sigma^2(X_{(i-1)\Delta}; \hat{\theta}_2)}(X_{i\Delta} - X_{(i-1)\Delta}) - \frac{\Delta}{2}\sum_{i=2}^{n_\Delta} \frac{\mu^2(X_{(i-1)\Delta}; \theta_1)}{\sigma^2(X_{(i-1)\Delta}; \hat{\theta}_2)}. \tag{17}$$

4. Asymptotic results

The asymptotic theory of a slightly different two-stage estimator in the multivariate case was obtained in Yoshida (1992) for models whose diffusion term is known up to a constant (matrix) factor, where both infill and long span asymptotics are employed both for the diffusion and drift parameter estimators. In this section we first derive the asymptotic theory for the same class of (scalar) models but only resort to long span asymptotics for the drift parameter asymptotic theory. We then investigate the asymptotic properties of the estimators proposed in Section 3 for model (1) whose diffusion function has a general form.

4.1. Scalar parameter in the diffusion function

To highlight the differences between our approach and Yoshida (1992), we first assume the data are generated according to the following stochastic differential equation:

$$dX_t = \mu(X_t; \theta_1^*)dt + \theta_2^* f(X_t)dB_t. \tag{18}$$

Denote θ_2^2 by τ and θ_2^{*2} by τ^*. Both $\mu(\cdot;\theta_1)$ and $f(\cdot)$ are time-homogeneous, $\mathcal{B}$-measurable functions on $\mathcal{D} = (l, u)$ with $-\infty \leq l < u \leq \infty$, where $\mathcal{B}$ is the σ-field generated by Borel sets on $\mathcal{D}$. τ is estimated by $\hat{\tau}$ defined by Eq. (7); θ_1 is estimated by $\hat{\theta}_1$, the maximizer of Eq. (8).

To prove consistency of $\hat{\tau}$ in a diffusion process with a constant diffusion term (i.e. $f(X_t) = 1$), Florens-Zmirou (1989) assumed $\Delta \to 0, T \to \infty$, and $\Delta^2 T \to 0$. The same set of assumptions were employed by Yoshida (1992) to deal with the diffusion process for more general, but still known, $f(X_t)$. In this paper, using the theory of Jacod (1994) and Barndorff-Nielsen and Shephard (2002), we show that the condition of an infinite time span of data (i.e. $T \to \infty$) is not needed to develop the asymptotic theory for $\hat{\tau}$, thereby extending the asymptotic results of Yoshida (1992) in a significant way.

We list the following conditions.

Assumption 1. Equation $[X]_t - \tau \int_0^t f^2(X_s)ds = 0$ has a unique solution at $\tau^* > 0\ \forall t > 0$.

Assumption 2. $\inf_{x\in J} f^2(x) > 0$, where J is a compact subset of the range of the process.

Assumption 3. $\int_0^t \mu^2(X_s; \theta_1)ds < \infty\ \forall t < \infty$.

Remark 4.1. Assumption 1 is an identification condition and Assumption 2 is a bounding positivity condition on the volatility function. Assumption 3 ensures weak convergence of the error process from the Euler approximation to the diffusion process (Jacod and Protter, 1998).

Theorem 4.1 (*Asymptotics of the Diffusion Parameter Estimate*). *Suppose Assumptions 1 and 2 hold, $\hat{\tau} \xrightarrow{p} \tau^*$ as $\Delta \to 0$. If, in addition, Assumption 3 holds,*

$$\Delta^{-1/2}(\hat{\tau} - \tau^*) \xrightarrow{d} \frac{\sqrt{2}\int_0^T \tau^* f^2(X_s)dW_s}{\int_0^T f^2(X_s)ds}$$

where W_t is a Brownian motion which is independent of X_t.

Remark 4.2. With a different estimate for τ, we substantially improve the results of Yoshida (1992), who derived asymptotic properties of the diffusion estimate assuming that $\Delta \to 0$ and $T \to \infty$, by only requiring in Theorem 4.1 that $\Delta \to 0$. Our result is not surprising and confirms the intuition that when the sampling interval goes to zero, the sample path within a finite time span, no matter how short, can perfectly reveal the quadratic variation of the process (see, for example, Merton (1980)) at least over that time span.

To establish the asymptotic properties of the drift parameter estimate, we follow Yoshida (1992) closely. In particular, we first list the following conditions.

Assumption 4. $\theta_1 \in \boldsymbol{\Theta}_1$ where the parameter space $\boldsymbol{\Theta}_1 \subset R^{K_1}$ is a compact set with $\theta_1^* \in Int(\boldsymbol{\Theta}_1)$.

Assumption 5. Both $\mu(\cdot;\theta_1)$ and $f(\cdot)$ functions are twice continuously differentiable. As a result, for any compact subset J of the range of the process, we have the following two conditions:

(i) (Lipschitz condition) There exists a constant L_1 so that

$$|\mu(x;\theta_1^*) - \mu(y;\theta_1^*)| + \theta_2^*|f(x) - f(y)| \leq L_1|x - y|,$$

for all x and y in J.

(ii) (Growth condition) There exists a constant L_2 so that

$$|\mu(x;\theta_1^*)| + \theta_2^*|f(x)| \leq L_2|1 + x|,$$

for all x and y in J.

Assumption 6. Define the scale measure of X_t by

$$s(x;\theta) = \exp\left(-2\int_c^x \frac{\mu(y;\theta_1)}{\tau f^2(y)}dy\right),$$

where c is a generic constant. We assume the following conditions hold

$$\int_c^u s(x;\theta)dx = \int_l^c s(x;\theta)dx = \infty,$$

and

$$\int_l^u \frac{1}{s(x;\theta)\tau f^2(x)}dx = A(\theta) < \infty.$$

Assumption 7. For arbitrary $p \geq 0$,

$$\sup_t E(|X_t|^p) < \infty.$$

Assumption 8. Define the following function

$$\theta_1 \to Y(\theta_1;\tau^*) = \int \frac{\mu(x,\theta_1)}{\tau^* f^2(x)}\left(\mu(x,\theta_1^*) - \frac{1}{2}\mu(x,\theta_1)\right)\pi_\theta(dx)$$

and assume function $Y(\cdot;\tau^*)$ has the unique maximum at $\theta_1 = \theta_1^*$, where π_θ is defined in Remark 4.4.

Assumption 9. For fixed θ_1, the derivatives $\partial^l\mu(x;\theta_1)/\partial x^l$ and $\partial^l f(x)/\partial x^l$ $(l = 1, 2)$ exist and they are continuous in x. For fixed x, $\partial^l\mu(x;\theta_1)/\partial\theta_1^l$ exist. Moreover,

$$|\partial^l\mu(x;\theta_1)/\partial x^l|, |\partial^l f(x)/\partial x^l|, |\partial^l\mu(x;\theta_1)/\partial\theta_1^l| \leq C(1 + |x|)^C,$$

for $l = 0, 1, 2$.

Assumption 10. The matrix

$$\Phi = \int \frac{\partial\mu(x;\theta_1^*)}{\partial\theta_1^\top}(\tau^* f^2(x))^{-1}\frac{\partial\mu(x;\theta_1^*)}{\partial\theta_1}\pi_\theta(dx) \tag{19}$$

is positive definite.

Remark 4.3. Under Assumption 5, these exists a solution process for the stochastic differential equation and the solution is unique.

Remark 4.4. Under Assumption 6, the process X_t is ergodic with an invariant probability measure that has density

$$\pi_\theta(x) = \frac{1}{A(\theta)s(x;\theta)\tau f^2(x)},$$

for $x \in (l, u)$ with respect to Lebesque measure on (l, u), where $A(\theta)$ and $s(x;\theta)$ are defined in Assumption 6. We further assume that $X_0 \sim \pi_{\theta^*}$ so that X_t is a stationary process with $X_t \sim \pi_{\theta^*}$.

Theorem 4.2 (*Asymptotics of the Drift Parameter Estimates*). *Let* $\hat{\theta}_1 = \text{argmax}_{\theta_1 \in \boldsymbol{\Theta}_1} T^{-1} \log \ell_{AIF}(\theta_1)$ *with* $\ell_{AIF}(\theta_1)$ *given by Eq. (8). Suppose Assumptions 1–10 hold,* $\hat{\theta}_1 \xrightarrow{p} \theta_1^*$ *as* $\Delta \to 0$ *and* $T \to \infty$. *If, in addition,* $\Delta^2 T \to 0$,

$$T^{1/2}(\hat{\theta}_1 - \theta_1^*) \xrightarrow{d} N(0, \Phi^{-1}),$$

where Φ *is given in Eq. (19).*

4.2. General diffusions

Now we consider the general case where Florens-Zmirou (1989) and Yoshida (1992) are not applicable. Suppose data are generated from the following stochastic differential equation

$$dX_t = \mu(X_t; \theta_1^*)dt + \sigma(X_t; \theta_2^*)dB_t, \tag{20}$$

where $\theta_1 \in \boldsymbol{\Theta}_1 \subset R^{K_1}$ and $\theta_2 \in \boldsymbol{\Theta}_2 \subset R^{K_2}$. Both $\mu(\cdot, \theta_1)$ and $\sigma(\cdot; \theta_2)$ are time-homogeneous, $\mathcal{B}$-measurable functions on $\mathcal{D} = (l, u)$ with $-\infty \le l < u \le \infty$, where $\mathcal{B}$ is the σ-field generated by Borel sets on $\mathcal{D}$. θ_2 is estimated by regressing (14) on (15), giving the extremum estimator in Box II; θ_1 is estimated by $\hat{\theta}_1$, defined by Eq. (17).

As in the scalar factor parameter case, we show that an infinite time span (i.e. $T \to \infty$) is not needed to develop the asymptotic theory for $\hat{\theta}_2$.

Some additional assumptions are required, given the nonlinear dependence of the diffusion $\sigma(X_t; \theta_2)$ on θ_2 **.** Also we have to modify some earlier Assumptions listed in Section 4.1.

Assumption 1′. The equation

$$[X]_t - \int_0^t \sigma^2(X_s; \theta_2)ds = \int_0^t \sigma^2(X_s; \theta_2^*)ds - \int_0^t \sigma^2(X_s; \theta_2)ds = 0 \tag{21}$$

has a unique solution at θ_2^*, $\forall t > 0$.

Assumption 2′. $\inf_{x \in J} \sigma^2(x; \theta_2^*) > 0$, where J is a compact subset of the range of the process.

Assumption 4′. $\theta_1 \in \boldsymbol{\Theta}_1$, $\theta_2 \in \Theta_2$, where parameter spaces $\boldsymbol{\Theta}_1 \subset R^{k_1}$ and $\Theta_2 \subset R^{k_2}$ are compact set with $\theta_1^* \in Int(\boldsymbol{\Theta}_1)$ and $\theta_2^* \in Int(\Theta_2)$.

Assumption 5′. Both $\mu(\cdot; \theta_1)$ and $\sigma(\cdot; \theta_2)$ functions are twice continuously differentiable. As a result, for any compact subset J of the range of the process, we have the following two conditions:

(i) (Lipschitz condition) There exists a constant L_1 so that

$$|\mu(x; \theta_1^*) - \mu(y; \theta_1^*)| + |\sigma(x; \theta_2^*) - \sigma(y; \theta_2^*)| \le L_1|x - y|,$$

for all x and y in J.

(ii) (Growth condition) There exists a constant L_2 so that

$$|\mu(x; \theta_1^*)| + |\sigma(x; \theta_2^*)| \le L_2|1 + x|,$$

for all x and y in J.

Assumption 6′. Define the scale measure of X_t by

$$s(x; \theta) = \exp\left(-2\int_c^x \frac{\mu(y; \theta_1)}{\sigma^2(y; \theta_2)}dy\right),$$

where c is a generic constant. We assume the following condition holds

$$\int_c^u s(x; \theta)dx = \int_l^c s(x; \theta)dx = \infty,$$

and

$$\int_l^u \frac{1}{s(x; \theta)\sigma^2(x; \theta_2)}dx = A(\theta) < \infty.$$

Assumption 8′. Define the following function

$$\theta_1 \to Y(\theta_1; \theta_2^*) = \int \frac{\mu(x, \theta_1)}{\sigma^2(x; \theta_2^*)}\left(\mu(x, \theta_1^*) - \frac{1}{2}\mu(x, \theta_1)\right)\pi_{\theta^*}(dx)$$

and assume $Y(\cdot; \theta_2^*)$ has the unique maximum at $\theta_1 = \theta_1^*$.

Assumption 9′. For fixed θ_1, the derivatives $\partial^l \mu(x; \theta_1)/\partial x^l$ and $\partial^l \sigma(x; \theta_2)/\partial x^l$ $(l = 1, 2)$ exist and they are continuous in x. For fixed x, $\partial^l \mu(x; \theta_1)/\partial \theta_1^l$ and $\partial^l \sigma(x; \theta_2)/\partial \theta_2^l$ exist. Moreover,

$$|\partial^l \mu(x; \theta_1)/\partial x^l|, |\partial^l \sigma(x; \theta_2)/\partial x^l|, |\partial^l \mu(x; \theta_1)/\partial \theta_1^l|, |\partial^l \sigma(x; \theta_2)/\partial \theta_2^l| \le C(1 + |x|)^C,$$

for $l = 0, 1, 2$.

Assumption 10′. The matrices

$$\Phi_1 = \int \frac{\partial \mu(x; \theta_1^*)}{\partial \theta_1}\sigma^{-2}(x; \theta_2^*)\frac{\partial \mu(x; \theta_1^*)}{\partial \theta_1'}\pi_\theta(dx)$$

and

$$\int_0^t \frac{\partial \sigma^2\left(X_s; \theta_2^*\right)}{\partial \theta_2}\frac{\partial \sigma^2\left(X_s; \theta_2^*\right)}{\partial \theta_2'}ds \tag{22}$$

are positive definite and $\int_0^t \sigma^4\left(X_s; \theta_2^*\right) ds > 0$ for all $t > 0$.

Theorem 4.3. *(Asymptotics of the Diffusion Parameter Estimate): Suppose Assumptions 1′–10′ hold. Then,* $\hat{\theta}_2 \xrightarrow{p} \theta_2^*$ *as* $\Delta \to 0$ *and*

$$\Delta^{-1/2}\left(\hat{\theta}_2 - \theta_2^*\right) \xrightarrow{d} \left[\sum_{k=1}^K \frac{\int_{(k-1)\frac{T}{K}}^{k\frac{T}{K}} \frac{\partial \sigma^2\left(X_s; \theta_2^*\right)}{\partial \theta_2}\frac{\partial \sigma^2\left(X_s; \theta_2^*\right)}{\partial \theta_2'}ds}{\int_{(k-1)\frac{T}{K}}^{k\frac{T}{K}} \sigma^4\left(X_s; \theta_2^*\right) ds}\right]^{-1} \times \left[\sum_{k=1}^K \frac{\sqrt{2}\int_{(k-1)\frac{T}{K}}^{k\frac{T}{K}} \frac{\partial \sigma^2\left(X_s; \theta_2^*\right)}{\partial \theta_2}\sigma^2\left(X_s; \theta_2^*\right) dW_s}{\int_{(k-1)\frac{T}{K}}^{k\frac{T}{K}} \sigma^4\left(X_s; \theta_2^*\right) ds}\right],$$

where W_t *is a Brownian motion which is independent of* X_t.

Theorem 4.4 (*Asymptotics of the Drift Parameter Estimate*). *Let* $\hat{\theta}_1 = \text{argmax}\, T^{-1} \log \ell_{AIF}(\theta_1)$ *with* $\ell_{AIF}(\theta_1)$ *given by Eq. (17). Suppose Assumptions* $1' - 10'$ *hold, then* $\hat{\theta}_1 \xrightarrow{p} \theta_1^*$ *as* $\Delta \to 0$ *and* $T \to \infty$. *If, in addition,* $\Delta^2 T \to 0$,

$$T^{1/2}(\hat{\theta}_1 - \theta_1^*) \xrightarrow{d} N(0, \Phi_1^{-1}),$$

where Φ_1 *is given in Eq. (22).*

5. Monte Carlo results

To examine the performance of the proposed procedure, we estimate the following model for short-term interest rates due to Chan et al. (1992, CKLS hereafter),

$$dX_t = \kappa(\mu - X_t)dt + \sigma X_t^\gamma dB_t, \tag{23}$$

with $\kappa = 0.6, \mu = 0.09, \sigma = 0.06, \gamma = 0.5$. We choose $\gamma = 0.5$ so that the true model becomes a CIR model which enables an exact data simulation. The parameters are estimated from 10 years of daily data (2500 observations). The experiment is replicated 1000 times to get the means and standard errors for each estimate. Two estimation methods are employed to estimate

Table 3
Simulation results under the CKLS model, $dX_t = \kappa(\mu - X_t)dt + \sigma X_t^{\gamma} dB_t$, based on 1000 samples of 2500 daily observations. True values for $\kappa, \mu, \sigma, \gamma$ are 0.6, 0.09, 0.06, 0.5. Mean and SD stand for the average and standard deviation across 1000 replications, respectively.

		AML	Two-Stage method									
			$K = 2$		$K = 10$		$K = 20$		$K = 50$		$K = 100$	
			Level	Log	Level	Log	Level	Log	Level	Log	Level	Log
γ	Mean	.4901	.5326	.5326	.4994	.4992	.4981	.4954	.4971	.4927	.4933	.4921
	SD	.1044	.5117	.5117	.1343	.1295	.1191	.1136	.1353	.1104	.1407	.1110
σ	Mean	.0604	.1562	.1562	.0628	.0628	.0612	.0612	.0589	.0603	.0588	.0602
	SD	.0157	.4754	.4754	.0235	.0235	.0177	.0177	.0190	.0167	.0202	.0169
κ	Mean	1.072	1.070	1.070	1.073	1.073	1.073	1.073	1.073	1.073	1.070	1.070
	SD	.5447	.5399	.5399	.5417	.5417	.5417	.5417	.5416	.5415	.5418	.5417
μ	Mean	.0901	.0902	.0902	.0902	.0902	.0902	.0902	.0902	.0902	.0902	.0902
	SD	.0097	.0094	.0094	.0094	.0094	.0094	.0094	.0094	.0094	.0094	.0094

the model: the approximate ML method of Aït-Sahalia (2002) and the proposed two-stage method. For the two-stage method, we implement it based on both levels and log levels, respectively. To use the two-stage methods, the number of subsamples has to be chosen. There is a trade-off between a large K and a small K. On the one hand, since K determines the number of observations used for the nonlinear regression, a larger K will generate more variation in $[X]_{k\frac{T}{K}} - [X]_{(k-1)\frac{T}{K}}$ across subsamples and hence provide more accurate estimation of the parameters in the diffusion function.[1] On the other hand, if K is too big, M_Δ will be too small for $\sum_{i=2}^{M_\Delta}(X_{(k-1)M_\Delta+i\Delta} - X_{(k-1)M_\Delta+(i-1)\Delta})^2$ to provide a good approximation to $[X]_{k\frac{T}{K}} - [X]_{(k-1)\frac{T}{K}}$, as Figs. 1–3 suggested. In the Monte Carlo study, we choose various values for K, namely 2, 10, 20, 50 and 100. These values correspond to 1250, 250, 125, 50 and 25 for M_Δ, respectively.

The simulation results are reported in Table 3. Several interesting results emerge from this table. First, the simulation results clearly indicate a trade-off between a large K and a small K. When K is 2, the variance is big for the two parameters in the diffusion function, indicating that the nonlinear least regression does not lead to accurate estimation to the diffusion parameters. The performance of the procedure improves substantially when $K \geq 10$. However, when K is large, a further increase in K fails to improve the finite sample performance. For example, for the two-stage method based on log-levels, the variances of the two diffusion parameters are slightly larger when $K = 100$ than those when $K = 50$. Second, the estimation of parameters in the drift function does not seem to be dependent on the diffusion parameters in any critical way. This is because the information matrix for the diffusion parameters is orthogonal to that for the drift parameters in the CKLS model. Thirdly, when a reasonable value for K is used, the quadratic variations are well estimated. Not surprisingly, therefore, our estimates are close to the approximate ML method of Aït-Sahalia (2002). Finally, we note that the two-stage method based on log levels has better finite sample performances than that based on levels, especially when K is large (and hence M_Δ is small), consistent with the finding in Barndorff-Nielsen and Shephard (2005).

6. Microstructure contamination

Direct application of the two-stage method proposed above requires that X_t be observed. This assumption may be too strong for ultra high frequency data because of the presence of various market microstructure effects which contaminate X_t with noise, producing bias and inconsistency in realized volatility estimates. As a consequence, our two-stage procedure has to be modified in order to produce consistent estimates of the parameters in the diffusion function.

Suppose the noise is orthogonal to X_t and independent and identically distributed over the grid $(\Delta, 2\Delta, \ldots, n_\Delta\Delta)$, i.e.

$$Y_{i\Delta} = X_{i\Delta} + e_{i\Delta}, \tag{24}$$

and

$$e_{i\Delta} \perp e_{j\Delta}, \quad \forall i \neq j \quad \text{and} \quad e_{i\Delta} \perp X_{i\Delta}, \quad \forall i, \tag{25}$$

with zero mean $E(e_{i\Delta}) = 0$ and finite variance $E(e_{i\Delta}^2) = \sigma^2, \forall i$. In the presence of market microstructure noise such as $e_{i\Delta}$, we observed $Y_{i\Delta}$ instead of $X_{i\Delta}$. While the pure noise assumption (25) lacks realism, as discussed in Phillips and Yu (2006), it is commonly adopted in the literature as it simplifies identification and econometric analysis – see, for example, Bandi and Russell (2005), Hansen and Lunde (2006), Barndorff-Nielsen et al. (2005) and Zhang et al. (2005). It is easy to show that as $\Delta \to 0$, the realized volatility quantity $\sum_{i=2}^{M_\Delta}(Y_{(k-1)M_\Delta+i\Delta} - Y_{(k-1)M_\Delta+(i-1)\Delta})^2$ diverges, making estimation in the first stage inconsistent.

To provide a consistent estimate of the quadratic variation differential $[X]_{k\frac{T}{K}} - [X]_{(k-1)\frac{T}{K}}$, one can use the realized kernel estimator of Barndorff-Nielsen et al. (2005), the two-scale estimator of Zhang et al. (2005), or the multi-scale estimator of Zhang (2006). While these procedures converge to the quadratic variation at slower rates than realized volatility in the absence of microstructure noise, they nonetheless provide consistent nonparametric estimators of the required quantity and can be employed to estimate the parameters in the diffusion function using the nonlinear regression, precisely as used above. Aït-Sahalia et al. (2006) generalized the two-scale procedure to cover time dependent noise and this generalization can also be adopted in the first step of our two-stage method in the same spirit. The asymptotic properties of the resulting estimator may be extracted along lines similar to those used here, with consequential changes in the rate of convergence. The results are obviously of interest and will provide an approach to continuous system estimation in the presence of noise. The details will be provided in subsequent work.

7. Concluding remarks

This paper proposes a two-stage method to estimate diffusion processes in a general form. In the first stage the realized volatility calculated from a sequence of split samples is regressed on the corresponding quadratic variation in order to estimate all the parameters in the diffusion function. Then, conditional on the resulting consistent estimate of the diffusion, the in-fill likelihood function approximation of the diffusion process can be readily constructed. The resulting discrete approximation produces estimates of all the parameters in the drift function. Monte Carlo simulations show that the finite sample performance of the proposed method is very satisfactory and as good

[1] We thank a referee for pointing this out to us.

as conventional maximum likelihood even when the discrete likelihood can be obtained. One advantage of the proposed method is that a larger scale optimization problem is decomposed into two smaller scale optimization problems. Although, like other extreme estimators, our method tends to over estimate the mean reversion parameter, κ, the numerical attractability of our method makes it an ideal initial estimate for the jackknife method of Phillips and Yu (2005) or the simulation-based methods of Phillips and Yu (in press-b) reduce the finite sample bias in κ.

There exist alternative two-step procedures to estimate diffusion processes. For example, in Bandi and Phillips (2007), nonparametric estimates of the drift and diffusion functions are matched with the parametric counterparts to provide consistent estimation of parameters in the drift and in the diffusion, respectively. One can certainly use our first step estimate (i.e. via realized volatility) in place of the diffusion estimate (i.e. via kernel functions) in the procedure of Bandi and Phillips (2007).

The approach can be readily extended to the multi-dimensional case. Both implementation and asymptotic theory only need trivial modifications. Since the method separates estimation of the drift and diffusion functions, it may be a desirable method to use when the drift but not the diffusion involves certain market microstructure features.

The two-step approach can be also adapted to deal with the following diffusion model mixed with jumps:

$$\mathrm{d}X_t = \mu(X_t;\theta_1)\mathrm{d}t + \sigma(X_t;\theta_2)\mathrm{d}B_t + f(X_t;\theta_3)\mathrm{d}Z_t^{\theta_3},$$

where $Z_t^{\theta_3}$ is a Lévy process with parameter θ_3. For this model realized volatility cannot be used to consistently estimate θ_2 as it does not converge to $\int \sigma^2(X_t;\theta_2)\mathrm{d}t$. However, empirical tripower variation does converge to $\int \sigma^2(X_t;\theta_2)\mathrm{d}t$. Barndorff-Nielsen and Shephard et al. (2006) obtained a central limit theorem for the tripower variation process when a diffusion process is mixed with finite activity jumps. Since the in-fill likelihood is readily available for jump-diffusion processes once the diffusion term is known, we can adopt the following two-step procedure to estimate the model: first, use tripower variation to estimate θ_2; then, maximize the approximated in-fill likelihood with respect to θ_1 and θ_3. The properties of the resulting estimator will be reported in future work.

Appendix

Proof of Theorem 4.1. It is known that all diffusion-type processes are semi-martingales (Prakasa Rao, 1999b). As a result, when $\Delta \to 0$,

$$[X_\Delta]_T \xrightarrow{p} [X]_T = \tau^* \int_0^T f^2(X_s)\mathrm{d}s,$$

where the convergence follows from the theory of quadratic variation for semi-martingales and the equality follows from Assumption 1.

By Assumption 3, we have

$$\hat{\tau} = \frac{[X_\Delta]_T}{\sum_{i=1}^{n_\Delta} f^2(X_{(i-1)\Delta})} \xrightarrow{p} \frac{[X]_T}{\int_0^T f^2(X_s)\mathrm{d}s} = \tau^*.$$

This proves the first part of Theorem 4.1.

Since X_t is a semi-martingale, by Ito's lemma for semi-martinagles (Prakasa Rao, 1999b) we have

$$X_T^2 = [X]_T + 2\int_0^T X_{s-}\mathrm{d}X_{s-}.$$

Following Theorem 1 of Barndorff-Nielsen and Shephard (2002) we have

$$\Delta^{-1/2}([X_\Delta]_T - [X]_T) \xrightarrow{d} \tau^*\sqrt{2}\int_0^T f^2(X_s)\mathrm{d}W_s, \tag{26}$$

where W_t is a Brownian motion which is independent of X_t. Hence,

$$\begin{aligned}\Delta^{-1/2}(\hat{\tau} - \tau^*) &= \Delta^{-1/2}\left(\frac{[X_\Delta]_T}{\sum_{i=1}^{n_\Delta} f^2(X_{(i-1)\Delta})} - \frac{[X]_T}{\int_0^T f^2(X_s)\mathrm{d}s}\right)\\ &= \Delta^{-1/2}\frac{1}{\int_0^T f^2(X_s)\mathrm{d}s}\left(\frac{\int_0^T f^2(X_s)\mathrm{d}s}{\sum_{i=1}^{n_\Delta} f^2(X_{(i-1)\Delta})}[X_\Delta]_T - [X]_T\right).\end{aligned} \tag{27}$$

By Assumption 3,

$$\frac{\sum_{i=1}^{n_\Delta} f^2(X_{(i-1)\Delta})}{\int_0^T f^2(X_s)\mathrm{d}s} \xrightarrow{p} 1.$$

By Slutsky's theorem, Eqs. (26) and (27) imply that

$$\Delta^{-1/2}(\hat{\tau} - \tau^*) \xrightarrow{d} \tau^*\frac{\sqrt{2}\int_0^T f^2(X_s)\mathrm{d}W_s}{\int_0^T f^2(X_s)\mathrm{d}s}. \tag{28}$$

This completes the proof of Theorem 4.1. ■

Proof of Theorem 4.2. Obviously, the proposed drift estimator is in the class of extremum estimators. Hence, one can prove consistency by checking sufficient conditions for extremum estimation problems. It is convenient here to check the conditions given in Newey and McFadden (1994, p.2121), namely, compactness, continuity, uniform convergence, and identifiability.

Compactness of $\boldsymbol{\Theta}$, continuity of $T^{-1}\log\ell_{AIF}(\theta_1;\hat{\tau})$ and the identification condition are assured by Assumption 1, Assumption 9 and Assumption 8, respectively. The uniform convergence of $T^{-1}\log\ell_{AIF}(\theta_1)$ to $Y(\theta_1,\tau^*)$ follows from Proposition 1, Lemma 1 and Lemma 2 in Yoshida (1992). Hence the first part of the theorem is proved.

To show asymptotic normality, we follow Yoshida (1992) by obtaining the weak convergence of the likelihood ratio random field,

$$Z_{\Delta,n_\Delta}(\tau,u) = \ell_{AIF}(\theta_1^* + T^{-1/2}u;\hat{\tau})/\ell_{AIF}(\theta_1^*;\hat{\tau}).$$

Under the listed conditions, Yoshida (1992) showed that

$$\begin{aligned}\log Z_{\Delta,n_\Delta}(\tau^*,u) &= u^\top T^{-1/2}\sum_{i=1}^{n_\Delta}\frac{\partial\mu(x;\theta_1^*)}{\partial\theta_1}\frac{1}{\tau^* f^2(X_{(i-1)\Delta})}\\ &\times\int_{(i-1)\Delta}^{i\Delta}\tau^* f(x)\mathrm{d}W_t - \frac{1}{2}u^\top\Phi u + \rho_{\Delta,n}(u),\end{aligned} \tag{29}$$

where $\rho_{\Delta,n}(u) \xrightarrow{p} 0$ and Φ is defined in Eq. (19).

From Theorem 4.1, we have, $\forall\eta > 0$, that there exists a Δ and a positive number c_1 such that

$$P(\Delta^{-1/2}(\hat{\tau} - \tau^*) > c_1) < \eta/2.$$

Let $\hat{\tau} = \tau^* + \Delta^{1/2}M$ and we have, $\forall\epsilon > 0$

$$\begin{aligned}&P(|\log Z_{\Delta,n_\Delta}(\hat{\tau},u) - \log Z_{\Delta,n_\Delta}(\tau^*,u)| > \epsilon)\\ &= P(\Delta^{-1/2}(\hat{\tau} - \tau^*) > c_1) + P(\sup_{|M|\le c_1}|\log Z_{\Delta,n}(\hat{\tau},u)\\ &\quad - \log Z_{\Delta,n_\Delta}(\tau^*,u)| > \epsilon)\\ &< \eta.\end{aligned} \tag{30}$$