High frequency data analysis of integrated continuous Itô semimartingales

Masaaki Fukasawa
Department of Mathematics, Osaka University

In this talk, we consider statistical inference from high frequency data

$$\{Y_{jh}; j = 0, 1, \dots, n\}, \quad h = 1/n \tag{1}$$

under a second-order continuous Itô semimartingale model

$$\begin{aligned} \mathrm{d}Y_t &= X_t \mathrm{d}t, \\ \mathrm{d}X_t &= V^0_t \mathrm{d}t + \sqrt{V_t}\mathrm{d}W_t, \end{aligned} \tag{2}$$

where V^0 and V are (unknown, unobservable) adapted processes and W is a standard Brownian motion. First we consider to estimate

$$\bar{V}[g] := \int_0^1 g(X_t, Y_t, t) V_t \mathrm{d}t$$

nonparametrically for a given continuous function g, and then, consider parametric estimation of θ when

$$V_t = \sigma(X_t, Y_t, t, \theta) \tag{3}$$

with a known function σ under the high frequency asymptotics $n \to \infty$.

Such a second-order model as (2) appeared at the very beginning of the history of stochastic differential equations; in the so-called Langevin dynamics, the equation of motion of small particles, introduced in 1907, is given by

$$\ddot{Y}_t = -\nabla q(Y_t) - \gamma \dot{Y}_t + \sigma \dot{W}_t,$$

where q is the potential of the system, γ is the coefficient of resistance and σ is a constant determined by γ, the temperature of the system and Boltzmann's constant. This is a special case of (2), with $X = \dot{Y}$, $V^0 = -\nabla q(Y) - \gamma X$ and $V = \sigma^2$. The Langevin model is a reformulation of Einstein's explanation for the Brownian motion of small particles in terms of atoms (or molecules) appeared in 1905. Note that the atomic theory was still not fully accepted at that time due to the lack of direct, experimental evidence. Perrin's experimental work in 1908-1913, which brought him the Nobel prize, was to test the model by estimating the parameter σ^2 from the trajectory of Y sampled at every 30 seconds. The estimate was consistent to the model and eventually, it was regarded as the first decisive evidence of the atomic nature of matter. See Bigg [2] and Newburgh et al. [13] for more details. Tracking the movement of one particle has been an effective experimental approach, especially in molecular biology. See e.g., Gittes and Schmidt [6].

If $V^0 = a(X, Y)$ with an affine function a and V is constant in (2), then Y is a Gaussian process. In such a framework, often under the name CAR (continuous-time auto-regressive) model, Bartlett [1], Brockwell et al. [3], Pandit and Wu [15] and Gloter [8] among others studied parametric estimation from low frequency data, that is,

$$\{Y_{jh}; j = 0, 1, 2, \dots, n\}, h > 0; \text{ fixed}, n \to \infty.$$

An extension to a model driven by a fractional Brownian motion is given by Tsai and Chan [19]. Ditlevsen and Sørensen [4] studied a case where X is a diffusion process under the low frequency asymptotics. Under the high frequency asymptotics (1) with $n \to \infty$, Gloter [7] studied parametric estimation of the diffusion coefficient (3). Gloter and Gobet [10] proved the LAMN property of the model. Gloter [9] and Nicolau [14] studied, respectively, parametric and nonparametric estimation problems under the mixed asymptotics

$$\{Y_{jh}; j = 0, 1, 2, \dots, n\}, h \to 0, \; nh \to \infty.$$

Pokern et al. [18] proposed a Bayesian approach. Papavasiliou et al. [16] and Pavliotis and Stuart [17] studied a case where (X, Y) is a diffusion with a homogenization structure under the mixed asymptotics.

There is a vast amount of literature for high frequency data analysis of first-order models, where

$$\{X_{jh}, j = 0, 1, \dots, n\}, \;\; h = 1/n$$

are given as data instead of (1). The pioneering work is Genon-Catalot and Jacod [5]. What play fundamental roles in studying first order models are that

$$\sum_{j=0}^{n-1} (X_{(j+1)h} - X_{jh})^2 \to \bar{V}[1] = \int_0^1 V_t \mathrm{d}t \tag{4}$$

in probability (Law of Large Numbers) and that

$$\sqrt{n} \left(\sum_{j=0}^{n-1} (X_{(j+1)h} - X_{jh})^2 - \bar{V}[1] \right) \to \mathcal{MN}\left(0, 2 \int_0^1 V_t^2 \mathrm{d}t\right)$$

stably in law (Central Limit Theorem) as $n = 1/h \to \infty$, where $\mathcal{MN}$ refers to a mixed normal distribution. In a second-order model where X_{jh} is not observed, the most natural proxy for X_{jh} would be the numerical derivative

$$X_j^h = \frac{1}{h}\int_{(j-1)h}^{jh} X_t \mathrm{d}t = \frac{1}{h}(Y_{jh} - Y_{(j-1)h}).$$

An interesting finding by Gloter [7] was that the approximation errors in the numerical derivatives are not negligible in the sense that

$$\sum_{j=0}^{n-1}(X_{j+1}^h - X_j^h)^2 \to \frac{2}{3}\bar{V}[1]$$

in probability. What makes this difference from (4) is the serial correlation of $X_{j+1}^h - X_j^h$ induced by the numerical differentiation. Gloter [7] showed that

$$\sqrt{n}\left(\frac{3}{2}\sum_{j=0}^{n-1}(X_{j+1}^h - X_j^h)^2 - \bar{V}[1]\right) \to \mathcal{MN}\left(0, \frac{9}{4}\int_0^1 V_t^2 \mathrm{d}t\right)$$

and based on this, constructed a $\sqrt{n}$ consistent estimator of θ. This estimator is however not asymptotically efficient in light of the LAMN result by Gloter and Gobet [10].

In this paper, we consider statistics of the form

$$\hat{V}_n^\kappa[1] := \sum_{i,j=1}^{n-1} \kappa(i-j)(X_{i+1}^h - X_i^h)(X_{j+1}^h - X_j^h)$$

or more generally,

$$\hat{V}_n^\kappa[g] := \sum_{i,j=1}^{n-1} \kappa(i-j) g(X_{i\wedge j}^h, Y_{(i\wedge j)h}, (i \wedge j)h)(X_{i+1}^h - X_i^h)(X_{j+1}^h - X_j^h) \tag{5}$$

for a given function g, where κ is a deterministic function on $\mathbb{Z}$. It is natural to expect that with a suitable choice of κ, the quadratic form $\hat{V}_n^\kappa[g]$ outperforms the diagonal forms used by Gloter [7] because it utilizes information of the serial correlation. In fact, when $V^0 = 0$ and V is constant, an asymptotically efficient estimator of $\bar{V}[1] = V$ turns out to be of the form. We show that the necessary and sufficient condition on κ for

$$\hat{V}_n^\kappa[g] \to \bar{V}[g]$$

to hold in probability as $n \to \infty$ is

$$\frac{2}{3}\kappa(0) + \frac{1}{3}\kappa(1) = 1.$$

Our main result is that

$$\sqrt{n}(\hat{V}^{\kappa}_{n}[g] - \bar{V}[g]) \to \mathcal{MN}\left(0, 2\int_0^1 g(X_t, Y_y, t)^2 V_t^2 \mathrm{d}t\right)$$

stably in law as $n \to \infty$, where

$$\kappa(k) = \sqrt{3}(\sqrt{3} - 2)^{|k|} \tag{6}$$

and that this choice of κ is optimal in the sense that it minimizes the asymptotic variance among the quadratic forms. We then apply this result to construct an asymptotically efficient estimator under parametric models.

The specification of κ as (6) is motivated by Whittle [20]'s approximation of covariance matrix, which has played an important role in the study of stationary time series. Our asymptotically efficient estimator is defined as the maximizer of

$$L_n(\theta) := -\frac{1}{2n}\sum_{j=0}^{n-1} \log \sigma(X^h_j, Y_{jh}, jh) - \frac{1}{2}\hat{V}^{\kappa}_{n}[\sigma(\cdot, \theta)^{-2}] \tag{7}$$

with κ defined by (6). This estimating function can be understood as the Whittle likelihood for high frequency data of the second-order model. Our data are not stationary, not Gaussian, and the model does not have the LAN property. This is seemingly the first study that shows the Whittle approximation can work beyond such classical frameworks.

References

[1] Bartlett, M.S. (1946): On the theoretical specification and sampling properties of autocorrelated time-series, *Suppl. J. Royal Stat. Soc.* 8, no. 1, 27-41.

[2] Bigg, C. (2008): Evident atoms: visuality in Jean Perrin's Brownian motion research, *Stud. Hist. Phil. Sci.* 39, 312-322.

[3] Brockwell, P.J., Davis, R.A. and Yang, Yu. (2007): Continuous-time Gaussian Autoregression, *Statist. Sinica* 17, 63-80.

[4] Ditlevsen, S. and Sørensen, M. (2004): Inference for observations of integrated diffusion processes, *Scand. J. Stat.* 31:417-429.

[5] Genon-Catalot, V. and Jacod, J. (1993): On the estimation of the diffusion coefficient for multidimensional diffusion processes, *Ann. Inst. H. Poincaré, Probab. Stat.* 29, 119-151.

[6] Gittes F, Schmidt C. (1998): Signals and Noise in Micromechanical Measurements. *Methods in Cell Biology* 55:129-156.

[7] Gloter, A. (2000): Discrete sampling of an integrated diffusion process and parameter estimation of the diffusion coefficient, *ESAIM Probab. Stat.* 4, 205 − 224.

[8] Gloter, A. (2001): Parameter estimation for a discrete sampling of an integrated Ornstein-Uhlenbeck process, *Statistics: A J. Theoret. Appl. Statist.* 35:3, 225-243.

[9] Gloter, A. (2006): Parameter Estimation for a discretely observed integrated diffusion process, *Scand. J. Stat.* 33:83-104.

[10] Gloter, A. and Gobet, E. (2008): LAMN property for hidden processes: The case of integrated diffusions, *Ann. Inst. H. Poincaré, Probab. Stat.* 44, no. 1, 104-128.

[11] Gobet, E. (2001): Local asymptotic mixed normality property for elliptic diffusion: a Malliavin calculus approach, *Bernoulli* 7(6), 899-912.

[12] Mattingly, J.C., Stuart, A.M. and Higham, D.J. (2002): Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise, *Stoch. Process. Appl.* 101, 185-232.

[13] Nerburgh, R., Peidle, J. and Reuckner, W. (2006): Einstein, Perrin, and the reality of atoms: 1905 revisited, *Am. J. Phys.* 74, 478.

[14] Nicolau, J. (2007): Nonparametric estimation of second-order stochastic differential equations, *Econometric Theory*, 23, 880-898.

[15] Pandit, S.M. and Wu, S.M. (1975): Unique estimates of the parameters of a continuous stationary stochastic process, *Biometrika* 62, 2, 497-501.

[16] Papavasiliou, A., Pavliotis, G.A. and Stuart, A.M. (2009): Maximum likelihood drift estimation for multiscale diffusions, *Stochastic Process. Appl.* 119, 3172-3210.

[17] Pavliotis, G.A. and Stuart, A.M. (2007): Parameter estimation for multiscale diffusions, *J. Stat. Phys.* 127, no. 4, 741-781.

[18] Pokern, Y., Stuart, A.M. and Wiberg, P. (2009): Parameter estimation for partially observed hypoelliptic diffusions, *J. R. Statist. Soc. B* 71, part 1, 49-73.

[19] Tsai, H. and Chan, K.S. (2005): Quasi-maximum likelihood estimation for a class of continuous-time long-memory processes, *J. Time Ser. Anal.* 26, no. 5, 691-713.

[20] Whittle, P. (1952): Estimation and information in time series analysis, *Scand. Aktuar.* 35, 48-60.