High Frequency Variability and Microstructure Bias

Adam Sykulski [*]
Dept of Mathematics,
Imperial College London
SW7 2AZ, London
ams03@ic.ac.uk

Sofia Olhede [†]
Dept of Statistical Science
University College London
WC1 E6BT, London
s.olhede@ucl.ac.uk

Grigorios Pavliotis [‡]
Dept of Mathematics,
Imperial College London
SW7 2AZ, London
g.pavliotis@ic.ac.uk

Abstract

This paper treats the multiscale estimation of integrated volatility of an Itô process immersed in high-frequency correlated noise. The multiscale structure of the problem is modelled explicitly, and the multiscale ratio is used to quantify energy contributions from the noise, estimated using the Whittle likelihood. This problem becomes more complicated as we allow the noise structure greater flexibility, and properties of the estimation is discussed via a simulation study.

1 Introduction

The estimation of properties of continuous time stochastic processes, whose observation is immersed in high frequency nuisance structure is required in many different fields of application, for example molecular biology and finance. Various methods have been proposed to alleviate bias introduced into the estimation from high frequency nuisance structure, see for example [1–4]. Commonly the model of the observed process is as the process of interest X_{t_i} superimposed with noise ε_{t_i}, or

$$Y_{t_i} = X_{t_i} + \varepsilon_{t_i}, \tag{1}$$

where Y_{t_i} is the observed process, X_{t_i} the unobserved component of interests, and ε_{t_i} is the microstructure noise effect. My model X_t, the process of interest with a suitable stochastic differential equation. For example, the Heston model is specified by ([5]):

$$dX_t = (\mu - \nu_t/2)\, dt + \sigma_t dB_t, \quad d\nu_t = \kappa\,(\alpha - \nu_t)\, dt + \gamma \nu_t^{1/2} dW_t, \tag{2}$$

where $\nu_t = \sigma_t^2$, and B_t and W_t are correlated 1-D Brownian motions.

Our main objective is to estimate the *integrated volatility*, $\langle X, X\rangle_T$ of the Itô process $\{X_t\}$, from the set of observations $\{Y_{t_i}\}$. Various different models have been proposed for the generation of X_{t_i} and ε_{t_i}. An outstanding problem is proposing more robust inference methods. [3] has relaxed the assumptions of [1], to include inference of processes with jumps. Another possible direction of development is to include more complicated noise structures, namely allowing for correlation between observations. The main issue with such relaxations, is that it becomes harder to separate energy due to the high frequency component from the process of interest.

Sykulski *et al.* [4] have proposed inference for multiscale processes based on the discrete Fourier transform. Fourier domain estimators have also been used for estimating noisy Itô processes, see [6], but the main innovation of Sykulski *et al.* is to present a theoretical framework for Harmonizable processes of interest, where an automatic procedure for estimating the nuisance structure is presented. The Whittle likelihood is used to estimate the energy level of the process of interest, as well

[*] www.ecs.soton.ac.uk/people/as07r
[†] www.homepages.ucl.ac.uk/~ucaksc0/
[‡] www.ma.ic.ac.uk/~pavl/

as the noise contamination. It was shown to perform well under various signal to noise scenarios, as well as path lengths, see [4].

The results of [4] or [1] are only appropriate when the noise is white. We therefore discuss possible extensions of the multiscale estimators to the case of more complicated market microstructure, and illustrate the performance of the estimator in various noise structure.

2 Multiscale Estimation

In the absence of noise a suitable estimator of the integrated volatility, $\langle X, X\rangle_T = \int_0^T \sigma_t^2\, dt$, can be specified from the quadratic variation of the process $\{Y_t\}$. In the presence of market microstructure noise this is no longer true and it is necessary to employ a different estimation procedure. For ease of exposition we denote the difference process $Z_{t_i} - Z_{t_{i-1}}$ by $U_{t_i}^{(Z)}$. The Loève spectrum of $U_{t_i}^{(Z)}$ will be denoted $S^{(Z)}(f_k, f_k)$, and we note that the observed quadratic variation can be rewritten as:

{e:naive}

$$\widehat{\langle X, X\rangle}_T^{(b)} = \sum_{i=0}^{N-1} \left(U_{t_i}^{(Y)}\right)^2 = \sum_{k=-N/2}^{N/2-1} \left|J^{(Y)}(f_k)\right|^2 \tag{3a}$$

$$\widehat{S}^{(Y)}(f_k, f_k) = \left|J^{(Y)}(f_k)\right|^2, \quad J^{(Y)}(f_k) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N-1} U_{t_j}^{(Y)} e^{-2\pi i t_j f_k}. \tag{3b}$$

with $f_k = \frac{k}{T}$. $\widehat{S}^{(Y)}(f_k, f_k)$ is the periodogram estimator, see [7], and normally has a single argument because the covariance of two fixed frequencies is asymptotically equivalent to zero for a stationary process. We note directly from [4] that the bias of the estimator is conveniently expressed in the Fourier domain by:

$$\mathrm{E}\left\{\widehat{\langle X, X\rangle}_T^{(b)}\right\} = \sum_{k=-N/2}^{N/2-1} S^{(X)}(f_k, f_k) + \sigma_\varepsilon^2 \sum_{k=-N/2}^{N/2-1} |2\sin(\pi f_k \Delta t)|^2 + O(N^\alpha) + O\left(N^{1-\alpha}\right). \tag{4}$$

The error terms follow from assumptions regarding the spectral properties of the noise process, and are detailed in [4], and this determines the value of α. It is clear from equation 4 that the influence of the noise increases for larger frequencies, and that the relative magnitude of $S^{(X)}(f_k, f_k)$ will determine the need for bias correction.

Sykulski *et al* proposed to measure the average energy of $U_t^{(X)}$ across frequencies, and determine the energy of $U_t^{(epsilon)}$, using the form of the white noise spectrum. Despite $U_t^{(X)}$ assumed harmonizable and not necessarily stationary, with appropriate assumptions regarding the spectral correlation of the process, it is appropriate to use the Whittle likelihood, see [8], to determine the energy of the two processes. Instead of using eqn 3a to estimate the quadratic variation a shrinkage estimator of $S^{(X)}(f_k, f_k)$ was proposed of

$$\widehat{S}^{(X)}(f_k, f_k; L_k) = L_k \widehat{S}^{(Y)}(f_k, f_k). \tag{5}$$

$0 \leq L_k \leq 1$ is referred to as the 'multiscale ratio' and its optimal form for perfect bias correction is given by:

$$L_k = \frac{S^{(X)}(f_k, f_k)}{S^{(X)}(f_k, f_k) + \sigma_\varepsilon^2 \left|2\sin(\pi f_k \Delta t)\right|^2}. \tag{6}$$

Of course this assumes perfect knowledge of $S^{(X)}(f_k, f_k)$ and is not a realizable estimator. Instead we worried about typical contributions of $S^{(X)}(f_k, f_k)$, and replaced the multiscale ratio by:

$$\overline{L}_k = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\varepsilon^2 \left|2\sin(\pi f_k \Delta t)\right|^2}. \tag{7}$$

The justification for this choice is discussed in Sykulski *et al*. We estimate the parameters of $\overline{L}_k$ by maximising a pseudo-likelihood namely the multiscale Whittle likelihood:

$$\ell(\sigma) = -\sum_{j=1}^{N/2-1} \log\left(S^{(X)}(f_j, f_j) + \sigma_\varepsilon^2 \left|2\sin(\pi f_j \Delta t)\right|^2\right) - \sum_{j=1}^{N/2-1} \frac{\widehat{S}^{(Y)}(f_j, f_j)}{S^{(X)}(f_j, f_j) + \sigma_\varepsilon^2 \left|2\sin(\pi f_j \Delta t)\right|^2}.$$

If $\left\{U_t^{(X)}\right\}$ is a stationary process, then the Whittle likelihood will approximate the time-domain likelihood of the sample, under suitable regularity conditions, see [9]. The bias corrected estimator of the integrated volatility then becomes

$$\widehat{\langle X,X\rangle}_T^{(m_1)} = \sum_{k=-N/2}^{N/2-1} \widehat{S}^{(X)}(f_k, f_k; \widehat{L}_k). \tag{8}$$

In Sykulski *et al.* it was shown that the estimates of σ_X^2 and σ_ϵ^2 produced bias corrected estimators with suitable properties. Unfortunately it is not always reasonable to model the high frequency structure as white, and so more subtle modelling needs to be applied.

3 Correlated Noise

A key issue is treating correlation in the error terms. A simple model would correspond to ϵ_{t_i} stationary but not white. Stationary processes can be conveniently represented in terms of aggregations of uncorrelated white processes, using the Wold decomposition theorem [10][p. 187]. We may therefore write the zero-mean process ϵ_{t_i} as

$$\epsilon_{t_i} = \sum_{j=0}^{\infty} \theta_{t_j} \eta_{t_i - t_j}, \tag{9}$$

where $\theta_{t_0} \equiv 1$, $\sum_j \theta_{t_j}^2 < \infty$, and $\{\eta_{t_n}\}$ satisfies $\mathrm{E}\left[\eta_{t_n}\eta_{t_m}\right] = \sigma_\eta^2 \delta_{n,m}$. Common practise would involve approximating the distribution by a finite number of elements in the sum, and thus replace ∞ by $q \in \mathbb{Z}$. We therefore model by

$$\epsilon_{t_i} = \eta_{t_i} + \sum_{k=1}^{q} \theta_{t_k} \eta_{t_i - t_k}, \tag{10}$$

and the spectral density of ϵ_{t_i} takes the form:

$$S^*(f; \boldsymbol{\theta}, \sigma_\eta^2) = \sigma_\eta^2 \left| 1 + \sum_{k=1}^{q} \theta_k e^{2i\pi f k} \right|^2. \tag{11}$$

In this case our spectral model for the noise changes to

$$S^{(\epsilon)}(f,f) = S^*(f; \boldsymbol{\theta}, \sigma_\eta^2) |2\sin(\pi f \Delta t)|^2. \tag{12}$$

Two possible methods now exist for treating the nuisance function of $S^*(f; \boldsymbol{\theta}, \sigma_\eta^2)$: we can use the method of Sykulski *et al.* directly without adjustment, assuming the variability of $S^*(f; \boldsymbol{\theta}, \sigma_\eta^2)$ to be moderate or we could adjust the methodology to encompass a parametric model for the noise, replacing $\widehat{\sigma}_\epsilon^2 |2\sin(\pi f \Delta t)|^2$ by $S^*(f; \widehat{\boldsymbol{\theta}}, \widehat{\sigma}_\eta^2) |2\sin(\pi f \Delta t)|^2$.

For a fixed and specified value of q, we may therefore estimate the parameters of the MA, using the Whittle likelihood, but where now $\sigma_\varepsilon^2 |2\sin(\pi f \Delta t)|^2$ is replaced by $S^*(f) |2\sin(\pi f \Delta t)|^2$. We thus get a multiscale likelihood given by

$$\begin{aligned}\ell(\sigma, \boldsymbol{\theta}) = &- \sum_{j=1}^{N/2-1} \log\left(S^{(X)}(f_j, f_j) + S^*(f_j; \boldsymbol{\theta}, \sigma_\eta^2) |2\sin(\pi f_j \Delta t)|^2 \right) \\ &- \sum_{j=1}^{N/2-1} \frac{\widehat{S}^{(Y)}(f_j, f_j)}{S^{(X)}(f_j, f_j) + S^*(f_j; \boldsymbol{\theta}, \sigma_\eta^2) |2\sin(\pi f_j \Delta t)|^2}.\end{aligned}$$

and the augmented multiscale ratio is given by

$$\overline{L}_k^{(a)} = \frac{\sigma_X^2}{\sigma_X^2 + S^*(f_j; \boldsymbol{\theta}, \sigma_\eta^2) |2\sin(\pi f_j \Delta t)|^2}. \tag{13}$$

If q is not assumed known, then model choice methods can also be applied to determine the value of q, such as applying the modified Akaike AIC [10][p. 287], and adding $2n(q+2)/[n-q-3]$ to minus two times the log multiscale likelihood, and minimizing this objective function. Some care must be applied as the Akaike AIC is known to overestimate the number of parameters, and BIC or some other model choice method may be applied. Of course the multiscale likelihood is not a true likelihood, and some care must be applied. Once we augment the estimation of σ_X^2 and σ_ε^2 with that of $\{\theta_{t_k}\}$, then we can estimate the noise spectrum and hence the multiscale ratio. This will yield an augmented estimator of the integrated volatility, that we denote by $\widehat{L}_k^{(a)}$. Our new estimator then takes the form

$$\widehat{\langle X,X\rangle}_T^{(a)} = \sum_{k=-N/2}^{N/2-1} \widehat{S}^{(X)}(f_k, f_k; \widehat{L}_k^{(a)}). \tag{14}$$

This form both takes the high frequency structure into account, and permits the high frequency structure to be more dynamic.

4 Examples

We investigate the simple case of

$$\varepsilon_{t_i} = \eta_{t_i} + \theta_1 \eta_{t_{i-1}}, \tag{15}$$

where then $S^*(f) = 1 + \theta_1^2 + 2\theta_1 \cos(2\pi f)$. Clearly it is of interest to investigate the effect of the variability of $S^*(f)$ on the multiscale estimation procedure. We start by illustrating the numerical effects of this on the existing estimation procedures, noting that of course in general it is inappropriate to use the methods without adjustment. As the range of θ goes from $(-1, 1)$ the spectral properties of the process changes quite markedly, from preferring high frequencies quite markedly to increasing energy at the low frequencies. If no adjustment is made to the proposed methods then one would anticipate overcompensating for the noise when θ is near minus one and under compensating when θ is near one. However, we remind the reader that the differencing has a very *strong* effect on the observed process. In effect the high frequency structure is so enhanced by the differencing that the minor effects of the MA parameters are negligible.

5 Conclusions

Acknowledgments

All acknowledgments go at the end of the paper.

References

[1] L. Zhang, P. A. Mykland, and Y. Ait-Sahalia, "A tale of two time scales: Determining integrated volatility with noisy high-frequency data", *J. Am. Stat. Assoc.*, vol. 100, pp. 1394–1411, 2005.

[2] G. A. Pavliotis and A. M. Stuart, "Parameter estimation for multiscale diffusions", *J. Stat. Phys.*, vol. 127, pp. 741–781, 2007.

[3] J. Fan and Y. Wang, "Multi-scale jump and volatility analysis for high-frequency financial data", *J. of the American Statistical Association*, vol. 102, pp. 1349–1362, 2007.

[4] A. Sykulski, S. C. Olhede, and G. Pavliotis, "Multiscale inference for high-frequency data", Tech. Rep. 290, Department of Statistical Science, University College London, 2008.

[5] Heston, "A closed form solution for options with stochastic volatility with applications to bond and currency options", *Review of Financial Studies*, vol. 6, pp. 327–343, 1993.

[6] M. E. Mancino and S. Sanfelici, "Robustness of fourier estimators of integrated volatility in the presence of microstructure noise", Tech. Rep., University of Firenze, 2006.

[7] D. Brillinger, *Time series: data analysis and theory*, Society for Industrial and Applied Mathematics, Philadelphia, USA, 2001.

[8] J. Beran, *Statistics for long-memory Processes*, Chapman and Hall, London, 1994.

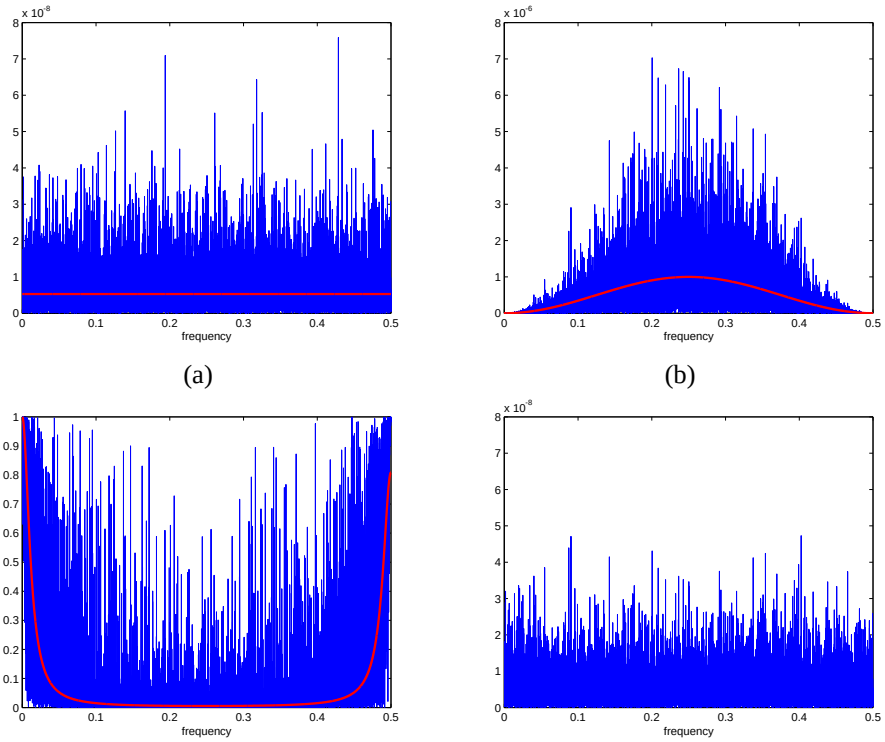


Figure 1: Estimating the frequency domain properties of the multiscale process. (a) True sample spectrum of

{fig0}

[9] K. O Dzhamparidze and A. M. Yaglom, "Spectrum parameter estimation in time series analysis", in *Developments in Statistics*, PR Krishnaiah, Ed., vol. 4, pp. 1–181. New York: Academic Press., 1983.

[10] Peter J. Brockwell and Richard A. Davis, *Time Series: Theory and Methods*, Springer, Berlin, Germany, 1991.