E C O N O M I C S B U L L E T I N

Integrated volatility measuring from unevenly sampled observations

Taro Kanatani
Graduate School of Economics, Kyoto University

Abstract

This paper derives the linear interpolation bias of realized volatility. To avoid the bias, the Fourier series estimator has been proposed by Malliavin and Mancino (2002). We examine the theoretical relationship between the Fourier estimator and realized volatility and show that the latter is the most efficient estimator in the class of the former.

The author would like to thank Kimio Morimune, Ryo Okui and seminar participants at Kyoto University for helpful comments and suggestions. This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology(MEXT), Grant−in−Aid for 21st Century COE Program "Interfaces for Advanced Economic Analysis".
Citation: Kanatani, Taro, (2004) "Integrated volatility measuring from unevenly sampled observations." *Economics Bulletin*, Vol. 3, No. 36 pp. 1−8
Submitted: July 25, 2004. **Accepted:** October 8, 2004.
URL: http://www.economicsbulletin.com/2004/volume3/EB−04C10020A.pdf

1 Introduction

Development of computer power and data recording systems allows us to use more high-frequency data than ever before. Such high-frequency data lend validity to a method based on quadratic variation formula, which is called *realized volatility* in the finance and econometrics literature. Because of the facility of handling, tick-by-tick data, which are usually unevenly sampled, are transformed into regularly spaced data through a certain interpolation. However, that interpolation method reduces the number of data and introduces the bias. Through Monte Carlo simulations, Barucci and Renò (2002) demonstrates that linear interpolation introduces a downward bias to realized volatilities. This paper theoretically derives the linear interpolation bias of realized volatility. To avoid these problems of interpolation methods, Malliavin and Mancino (2002) proposed an estimator based on Fourier series analysis that is well suited for unevenly sampled data. In this paper, we derive a theoretical relationship between the Fourier series estimator and realized volatility. The latter is proved to be the most efficient estimator in the class of the former. We confirm our theory through a Monte Carlo simulation.

This study specifically addresses the following situation. Let p_t be a logarithmic asset price that is generated by diffusion:

$$dp(t) = \sigma(t)dW(t),\ 0 \leq t \leq T, \tag{1}$$

where $W(t)$ is a standard Brownian Motion and $\sigma(t)$ is a random time dependent function. That diffusion is observed at $(N+1)$ irregular time points $\{t_i\}_{i=0}^{N}$. We assume that every time difference (duration) is sufficiently small: $\lim_{N\to\infty} \sup_{i\geq 1} (t_i - t_{i-1}) = 0$. For purposes of simplification, we set the drift of diffusion as 0. This simplification is acceptable not only because it means an efficient market in financial economics, but also because, mathematically, the martingale component swamps the predictable portion over short time intervals. In such a situation, we study the nonparametric estimators of integrated volatility $\int_0^T \sigma^2(t)dt$. Because we make no hypothesis on the structure of the underlying probability space Ω, we can construct an auxiliary probability space X where we consider $\sigma(t)$ as a deterministic function, see Malliavin and Mancino (2002).

Throughout this paper, E denotes the expectation on the probability space X.

2 Realized volatility from evenly spaced observations

In this section, we examine realized volatility from evenly spaced data. Unevenly sampled raw data are converted into evenly spaced data through interpolation. We consider two interpolation methods for converting N raw data, $\{p(t_i)\}_{i=0}^{N}$, to m evenly spaced data, $\{q(jT/m)\}_{j=0}^{m}$:

$$q\left(\frac{jT}{m}\right) = \begin{cases} (1-\rho_j)\,p\left(t_j^-\right) + \rho_j p\left(t_j^+\right) & \text{linear interpolation} \\ p\left(t_j^-\right) & \text{previous-tick interpolation} \end{cases} \tag{2}$$

where

$$\begin{aligned} \rho_j &= \frac{(jT/m) - t_j^-}{t_j^+ - t_j^-}, \\ t_j^- &= \max\{t_i : t_i \le jT/m\}, \\ t_j^+ &= \min\{t_i : t_i \ge jT/m\}, \end{aligned}$$

and where $\max A$ and $\min A$ denote maximum and minimum elements of A, respectively.

Using the evenly spaced data series $\{q(jT/m)\}_{j=0}^{m}$, the volatility is measured by the following estimator.

$$\hat{\sigma}^2(m) = \sum_{j=1}^{m}\left(q\left(\frac{jT}{m}\right) - q\left(\frac{(j-1)T}{m}\right)\right)^2. \tag{3}$$

Whereas Barucci and Renò (2002) found through Monte Carlo simulation that linear interpolation procedures introduce bias into realized volatility (3), we derive the theoretical linear interpolation bias as

$$-\frac{m}{T}\sum_{j=1}^{m}\left\{\rho_{j-1}(1-\rho_{j-1})\left(t_{j-1}^+ - t_{j-1}^-\right) + \rho_j(1-\rho_j)\left(t_j^+ - t_j^-\right)\right\}\int_{(j-1)T/m}^{jT/m}\sigma^2(t)\,dt. \tag{4}$$

See Appendix A for derivation of (4). Note that the downward bias (4) is more pronounced: (a) when the time window $[0,T]$ is divided more finely (m is large); (b) when the interpolated time point is far from the observed time points ($|\rho_j - 1/2|$ is small); (c) in coarsely-sampled periods ($t_j^+ - t_j^-$ is large); or (d) in the volatile period ($\sigma^2(t)$ is large).

In the case of previous-tick interpolation, the realized volatility (3) is unbiased if the diffusion is observed at $t = 0$ and $t = T$ (if $t_0 = 0$ and $t_N = T$).

3 Estimators using raw data

Malliavin and Mancino (2002) proposed a method based on Fourier series to use unevenly sampled data. In this section, we normalized the time window $[0, T]$ to $[0, 2\pi]$. The Fourier estimator of integrated volatility $\int_0^T \sigma^2(t)dt$ is given as

$$\hat{\sigma}_F^2 = \frac{\pi^2}{K} \sum_{k=1}^{K} (a_k^2(dp) + b_k^2(dp)), \tag{5}$$

where

$$a_k(dp) = \frac{1}{\pi} \int_0^{2\pi} \cos(kt) dp(t), \tag{6}$$

$$b_k(dp) = \frac{1}{\pi} \int_0^{2\pi} \sin(kt) dp(t), \tag{7}$$

and K is a large integer. In practice, we compute the integrals (6) through integration by parts, as

$$\begin{aligned} a_k(dp) &= \frac{1}{\pi} \int_0^{2\pi} \cos(kt) dp(t) = \frac{p(2\pi) - p(0)}{\pi} + \frac{1}{\pi} \int_0^{2\pi} \sin(kt) p(t) dt \\ &\approx \frac{p(2\pi) - p(0)}{\pi} + \frac{1}{\pi} \sum_{i=0}^{N-1} \left[\cos(kt_i) - \cos(kt_{i+1})\right] p(t_i) \end{aligned} \tag{8}$$

because the piecewise constant is valid under assumption $\lim_{N\to\infty} \sup_{i\geq 1} (t_i - t_{i-1}) = 0$. Similarly, we approximate (7) by

$$b_k(dp) \approx \frac{1}{\pi} \sum_{i=0}^{N-1} \left[\sin(kt_i) - \sin(kt_{i+1})\right] p(t_i). \tag{9}$$

Another method using unevenly sampled observations $\{p(t_i)\}_{i=0}^N$ is an estimator based on the quadratic variation formula:

$$\hat{\sigma}_R^2 = \sum_{i=1}^{N} \{\Delta p(t_i)\}^2, \tag{10}$$

where $\Delta p(t_i) = p(t_i) - p(t_{i-1})$. To distinguish (10) from (3), we refer to (10) as *raw data realized volatility*. The Fourier estimator (5) can be rewritten as

$$\hat{\sigma}_F^2 = \hat{\sigma}_R^2 + \sum_{i \neq j} \Delta p(t_i) \Delta p(t_j) w_{ij}, \tag{11}$$

where

$$w_{ij} = \frac{\sin\frac{(K+1)(t_i-t_j)}{2}\cos\frac{K(t_i-t_j)}{2}}{K\sin\frac{(t_i-t_j)}{2}}. \tag{12}$$

See Appendix B for the derivation. (11) and (12) imply that as $K \to \infty$, $\hat{\sigma}_F^2 \to \hat{\sigma}_R^2$. Because the second parts of (11) are uncorrelated to $\hat{\sigma}_R^2$,

$$\begin{aligned} V\left(\hat{\sigma}_F^2\right) &= V\left(\hat{\sigma}_R^2\right) + V\left(\sum_{i\neq j}\Delta p\left(t_i\right)\Delta p\left(t_j\right)w_{ij}\right) \\ &= V\left(\hat{\sigma}_R^2\right) + \sum_{i\neq j}\left\{\int_{t_{i-1}}^{t_i}\sigma^2\left(t\right)dt\right\}\left\{\int_{t_{j-1}}^{t_j}\sigma^2\left(t\right)dt\right\}w_{ij}^2 > V\left(\hat{\sigma}_R^2\right). \end{aligned} \tag{13}$$

In other words, as $K \uparrow \infty$, $V\left(\hat{\sigma}_F^2\right) \downarrow V\left(\hat{\sigma}_R^2\right)$. That is to say, (10) is the most efficient estimator in the class of (5).

4 Monte Carlo study

We follow the Monte Carlo design of Barucci and Renò (2002) with little modification: we generate a proxy for continuous observation by discretizing the following stochastic differential equations with a time step of one second:

$$\begin{aligned} d\log\sigma^2\left(t\right) &= -k\log\sigma^2\left(t\right)dt + \gamma dW_\sigma\left(t\right) \\ dp(t) &= \sigma(t)dW_p(t),\ 0 \le t \le T, \end{aligned}$$

where W_σ and W_p are mutually independent standard Brownian motions, $k = 0.01, \gamma = 0.1$, and $T = 60 \times 60 \times 24$ seconds (s). Time differences are drawn from an exponential distribution with a mean of 45 s.[1] We compare the performances of estimators (3), (5), and (10). In calculations of (3), we set $m = 144,\ 288$, and 720, corresponding to so-called daily realized volatility based on 10-min, 5-min, and 2-min returns. Each return is computed by two interpolation methods in (2). In (5), we set $K = 10,\ 50,\ 100,\ 500$, and $[N/2]$ where $[\cdot]$ denotes a Gaussian symbol.[2] In our simulation, $[N/2]$ is expected to be around $60 \times 60 \times 24 \div (45 \times 2) = 960$.

[1] Although each duration is independent in our simulation, our method requires no assumption except that $\sup_{i\geq 1}\left(t_i - t_{i-1}\right)$ is small. See Engle and Russell (1998) for the autoregressive time duration models. See e.g., Aït-Sahalia and Mykland (2003) for an example of the exponentially distributed duration.

[2] $[N/2]$ is the so-called Nyquist frequency if observations are sampled evenly.

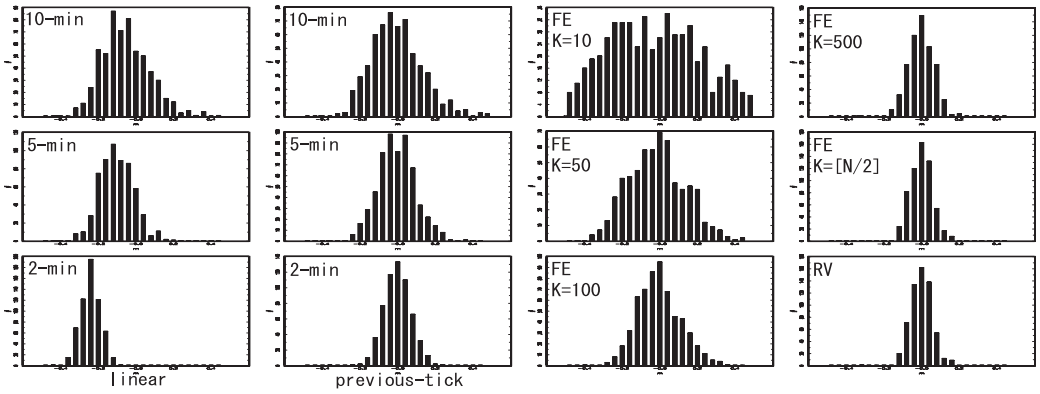

Figure 1: Distribution of (14). 10-min, 5-min, and 2-min denote the estimators (3) with $m = 144$, 288, and 720, respectively, through linear and previous-tick interpolation. FE signifies the Fourier estimator (5) with $K = 10$, 50, 100, 500, and $[N/2]$. RV denotes the raw data realized volatility (10). The distribution is computed with 600 'daily' replications.

We performed 600 replications. Figure 1 shows the distributions of normalized errors

$$\frac{\hat{\sigma}^2(m) - \int_0^T \sigma^2(t)dt}{\int_0^T \sigma^2(t)dt}, \quad \frac{\hat{\sigma}_F^2 - \int_0^T \sigma^2(t)dt}{\int_0^T \sigma^2(t)dt}, \text{ and } \quad \frac{\hat{\sigma}_R^2 - \int_0^T \sigma^2(t)dt}{\int_0^T \sigma^2(t)dt}. \tag{14}$$

Table 1 reports the means and standard deviations of (14) from that set of 600 replications. Increasing the number of partitions m, we can reduce the variance of realized volatility. However, as stated in (4), in the case of using linear interpolation, the downward bias increases. In contrast, the realized volatility is unbiased when using previous-tick interpolation. As stated in (13), Figure 1 and Table 1 show that as $K \uparrow \infty$, $V(\hat{\sigma}_F^2) \downarrow V(\hat{\sigma}_R^2)$. Table 2 compares means of the theoretical linear interpolation bias (4) and measurement error of $\hat{\sigma}^2(m) - \int_0^T \sigma^2(t)dt$ from 600 replications. Both means are approximately consistent. We can verify the validity of (4) by these results.

5 Concluding remarks

This study derived the linear interpolation bias of realized volatility. Results indicate that linear interpolation should not be used as the preparation for realized volatility calculations. The theoretical relationship between the

Table 1: Means and standard deviations of (14) from 600 'daily' replications. Standard deviations are given in parentheses. 10-min, 5-min, and 2-min denote the estimators (3) with $m = 144$, 288, and 720, respectively, through two different interpolations in (2). Fourier estimators are computed with five different Ks. Raw data realized volatility denotes the estimator (10).

linear	10-min	-0.04734	(0.12293)
	5-min	-0.09763	(0.08937)
	2-min	-0.23911	(0.05152)
previous-tick	10-min	0.00109	(0.13139)
	5-min	-0.00092	(0.09864)
	2-min	-0.00017	(0.07086)
Fourier estimator	$K = 10$	0.00107	(0.33051)
	$K = 50$	-0.01371	(0.15253)
	$K = 100$	-0.00341	(0.11409)
	$K = 500$	-0.00252	(0.06538)
	$K = [N/2]$	-0.00055	(0.05730)
raw data realized volatility		-0.00094	(0.05460)

Table 2: Means of measurement errors of realized volatility (3) through linear interpolation procedures and the theoretical linear interpolation bias (4) from 600 replications. Standard deviations are given in parentheses.

	$\hat{\sigma}^2(m) - \int_0^T \sigma^2(t)dt$		bias (4)	
10-min	-5291.56	(13682.3)	-5436.57	(478.465)
5-min	-10858.1	(9953.48)	-10915.0	(720.160)
2-min	-26675.4	(5907.82)	-27360.9	(1464.48)

Fourier series estimator proposed by Malliavin and Mancino (2002) and raw data realized volatility implies that the latter is the most efficient estimator in the class of the former.

A Linear interpolation bias

Using

$$q\left(\frac{jT}{m}\right) = (1-\rho_j)\, p\left(t_j^-\right) + \rho_j p\left(t_j^+\right) = (1-\rho_j)\int_0^{t_j^-} dp\,(t) + \rho_j \int_0^{t_j^+} dp\,(t)\,,$$

we get

$$\begin{aligned}
&E\left[q\left(\frac{jT}{m}\right) - q\left(\frac{(j-1)\,T}{m}\right)\right]^2 \\
&= \left[\rho_j^2 \int_0^{t_j^+} \sigma^2\,(t)\,dt + \left(1-\rho_j^2\right)\int_0^{t_j^-} \sigma^2\,(t)\,dt\right] \\
&- \left[\left(1-(1-\rho_{j-1})^2\right)\int_0^{t_{j-1}^+} \sigma^2\,(t)\,dt + \left(1-\rho_{j-1}^2\right)\int_0^{t_{j-1}^-} \sigma^2\,(t)\,dt\right].
\end{aligned}$$

Then calculating

$$\begin{aligned}
&E\left[\hat{\sigma}^2\,(m)\right] - \int_0^T \sigma^2\,(t)\,dt \\
&= \sum_{j=1}^{m} E\left[q\left(\frac{jT}{m}\right) - q\left(\frac{(j-1)\,T}{m}\right)\right]^2 - \int_{(j-1)T/m}^{jT/m} \sigma^2\,(t)\,dt,
\end{aligned}$$

we obtain (4).

B Fourier estimator and realized volatility

(8) and (9) are rewritten respectively as

$$a_k(dp) = \frac{1}{\pi}\sum_{i=1}^{N} \cos kt_i \Delta p(t_i) \tag{15}$$

$$b_k(dp) = \frac{1}{\pi}\sum_{i=1}^{N} \sin kt_i \Delta p(t_i). \tag{16}$$

Using (15), (16), and the addition theorem,

$$\begin{aligned}
\hat{\sigma}_F^2 &= \frac{\pi^2}{K}\sum_{k=1}^{K}(a_k^2(dp)+b_k^2(dp_i)) \\
&= \frac{1}{K}\sum_{k=1}^{K}\left(\left\{\sum_{i=1}^{N}\cos kt_i\Delta p(t_i)\right\}^2+\left\{\sum_{i=1}^{N}\sin kt_i\Delta p(t_i)\right\}^2\right) \\
&= \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N}\Delta p\left(t_i\right)\Delta p\left(t_j\right)\left\{\cos kt_i\cos kt_j+\sin kt_i\sin kt_j\right\} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\Delta p\left(t_i\right)\Delta p\left(t_j\right)\left\{\frac{\sum_{k=1}^{K}\cos k\left(t_i-t_j\right)}{K}\right\}.
\end{aligned}$$

Because

$$\sum_{k=1}^{K}\cos kx=\begin{cases} K & \text{if } x=0 \\ \frac{\sin\frac{(K+1)x}{2}\cos\frac{Kx}{2}}{\sin\frac{x}{2}} & \text{otherwise,}\end{cases}$$

we obtain(11).

References

Aït-Sahalia, Y. and P. A. Mykland (2003). The effects of random and discrete sampling when estimating continuous-time diffusions. *Econometrica 71*, 483–549.

Barucci, E. and R. Renò (2002). On measuring volatility of diffusion processes with high frequency data. *Economics Letters 74*, 371–378.

Engle, R. F. and J. R. Russell (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica 66*, 1127–1162.

Malliavin, P. and M. E. Mancino (2002). Fourier series method for measurement of multivariate volatilities. *Finance and Stochastics 6*, 49–61.