

Unbiased covariance estimation with interpolated data*

Taro Kanatani[†]

JSPS Research Fellow, Institute of Economic Research, Kyoto University

Yoshida Honmachi, Sakyo-Ku, Kyoto 6068501, Japan

Roberto Renò[‡]

Dipartimento di Economia Politica, Università di Siena

Piazza S. Francesco 7, 53100, Siena, Italy

April 4, 2007

Abstract

We study covariance estimation when compelled to use evenly spaced data which have already been manipulated by previous-tick interpolation. We propose an unbiased covariance estimator, which is designed to correct for the two biases arising because of the interpolation: non-synchronous trading and zero-return bias. We show how these sources make usual realized covariance estimators biased, and that the traditional lead-lag modification does not correct these biases completely. The proposed estimator is also proved to be consistent with the Hayashi and Yoshida (2005)'s unbiased estimator under extremely high frequency situation. We illustrate the potential advantages of the method with both simulated and actual data.

Keywords: Realized covariance; Previous tick interpolation; Epps effect; Nonsynchronous trading; Bias-correction

JEL Classification: C14; C32; C63

*Taro Kanatani's research is supported by Grant-in-Aid for JSPS fellows. Data were purchased from Nikkei Media Marketing, Inc. by the Institute of Statistical Mathematics. This study was carried out under the ISM Cooperative Research Program No. 2006-ISMCRP-1027, and under which the use of data in this study is authorized.

[†]E-mail: kanatani@kier.kyoto-u.ac.jp

[‡]E-mail: reno@unisi.it

1 Introduction

Estimating covariances is important in many applications. Given two discretely observed time series, if they are observed at the same time instants there is no difference in estimating covariance or variances, since $4 \cdot Cov(X, Y) = Var(X + Y) - Var(X - Y)$. The problem of estimating covariances arises when the two time series are not observed synchronously. Such ‘nonsynchronous observation’ problem has been solved by the estimator proposed in Hayashi and Yoshida (2005) (henceforth HY). However, there are many situations in which the time series are observed at different instants, but interpolated to new time series, which carry less information than the original ones. Under these situations, the estimator developed by HY cannot be implemented, since it needs not only the observations, but also the time instants of both time series. The main contribution of this paper is to present a methodology for measuring the covariance of two discretely observed time series, when they are first observed discretely at random points in time, then interpolated using previous-tick interpolation to get an evenly spaced time series.

This case is of particular importance for estimating the covariance of financial assets, which is widely called co-volatility or cross-volatility in the financial econometrics literature. The most fundamental examples arise from the literature on intraday data. Financial assets (stocks, bonds, commodities and so on) trade with very different intensities, ranging from less than a second for the most liquid, to several hours for the less liquid. In this situation, it is typical to analyze data which have been interpolated at a given frequency, e.g. one minute. Many data vendors distribute data in this form.

The recent interest of financial econometrics in high frequency data led to the flourishing of *realized* estimators for high frequency data, including realized covariance (Andersen et al., 2001, 2003; Barndorff-Nielsen and Shephard, 2004) and many refinements, see e.g. Griffin and Oomen (2006). Since the empirical study by Epps (1979), it is well known that the bias comes from non-synchronicity of the data, and different solutions have been proposed to correct for the non-synchronicity problem (Scholes and Williams, 1977; Cohen et al., 1983; Lo and MacKinlay, 1990). The bias can be prevailing in the intraday domain, since realized covariance is more and more biased toward zero as the sampling frequency increases.

We analyze the sources of the bias, showing that these can be divided into two: non-synchronous bias and zero return bias. We show that traditional methods cannot correct both, and we propose a method to handle these two biases. We then test the methodology on simulated and actual data. Simulated data help in confirming the theory, and to provide an order of magnitude of the time scales at which the bias become prominent. In our

application, we compute bias corrected covariance estimates with intraday stock prices, comparing the results with those obtained from competing estimators. Our conclusion is that under high frequency situation, the proposed estimator should be selected as the most reliable covariance measure.

The rest of the paper is organized as follows. In Section 2, we introduce the data generating process we assume throughout the paper and take a look at the realized covariance matrix. In Section 3, we illustrate the nonsynchronous bias of realized covariance and introduce traditional lead and lag modification. In Section 4, we show an example in which the traditional method is not enough and propose a new bias-corrected estimator. In Section 5, we confirm our theory through a Monte Carlo study. We present an application to financial data in Section 6. The final Section is devoted to concluding remarks.

2 Realized covariance

We consider a multi-dimensional stochastic process, e.g. representing a logarithmic asset price vector. Without loss of generality, we limit our discussion to the two-dimensional case.

Assumption 1 $p(t)$ is an \mathbb{R}^2 -valued stochastic process in $[0, T]$, driven by

$$\begin{pmatrix} dp_1(t) \\ dp_2(t) \end{pmatrix} = \begin{pmatrix} \sigma_{11}(t) & \sigma_{12}(t) \\ 0 & \sigma_{22}(t) \end{pmatrix} \begin{pmatrix} dW_1(t) \\ dW_2(t) \end{pmatrix}, \quad (2.1)$$

where $(W_1(t), W_2(t))$ is a standard two-dimensional Brownian motion, and $\sigma_{ij}(t)$ is adapted, measurable and bounded stochastic processes such that a unique solution of the SDE (2.1) exists with the initial condition $(p_1(0), p_2(0)) \in \mathbb{R}^2$.

The zero-drift assumption is allowable, not only because it means an efficient market in financial economics, but also because, mathematically, the martingale component swamps the drift over short time intervals. The time-varying covariance matrix $\Omega(t)$ is defined as:

$$\begin{pmatrix} \Omega_{11}(t) & \Omega_{12}(t) \\ \Omega_{12}(t) & \Omega_{22}(t) \end{pmatrix} \equiv \begin{pmatrix} \sigma_{11}(t)^2 + \sigma_{12}(t)^2 & \sigma_{12}(t)\sigma_{22}(t) \\ \sigma_{12}(t)\sigma_{22}(t) & \sigma_{22}(t)^2 \end{pmatrix}, \quad (2.2)$$

The estimation target is the value of the integrated covariance matrix over a fixed time interval $[0, T]$:

$$\int_0^T \Omega_{12}(t) dt.$$

For estimating this matrix, the following result is well known:

Proposition 2 *If all the components of $p(t)$ are observed at the same time instants,*

$$0 = t_0 < t_1 < \dots < t_N \leq T,$$

and

$$\lim_{N \rightarrow \infty} \max_n (t_n - t_{n-1}) = 0$$

then, under Assumption 1, we have

$$p - \lim_{N \rightarrow \infty} \sum_{n=1}^N (p(t_n) - p(t_{n-1}))(p(t_n) - p(t_{n-1}))' = \int_0^T \Omega(t) dt.$$

This is the theoretical basis of realized covariance, see e. g. Barndorff-Nielsen and Shephard (2004). Proposition 2 is based on the idea of synchronous and continuous record. However, actual financial observations are made at non-synchronous time instants.

Definition 3 *We define a partition \mathcal{P}_N of the interval $[0, T]$ as the set of $N+1$ increasing time instants covering the whole interval:*

$$\mathcal{P}_N = \{t_0, t_1, \dots, t_N : 0 = t_0 < t_1 < \dots < t_N = T\}$$

Two partitions $\mathcal{P}_{N_1}, \mathcal{P}_{N_2}$ are said completely asynchronous if $\mathcal{P}_{N_1} \cap \mathcal{P}_{N_2} = \{0, T\}$. A partition \mathcal{P}_M is said evenly spaced if $t_m - t_{m-1} = \frac{T}{M}$, $m = 1, \dots, M$

We consider the situation where each component of $p(t)$ is observed at a partition \mathcal{P}_{N_i} , $i = 1, 2$, and data are successively interpolated to an evenly spaced grid, according to the so-called *previous tick* interpolation scheme. The previous-tick interpolation is defined as follows. Define:

$$t_m^i = \max \left\{ t \in \mathcal{P}_{N_i} : t \leq \frac{mT}{M} \right\}, \quad i = 1, 2. \quad (2.3)$$

The interpolated time series are defined as:

$$q_i(m) = p_i(t_m^i), \quad m = 0, \dots, M, \quad i = 1, 2, \quad (2.4)$$

that is, we denote by p the original time series and by q the interpolated time series. To sum up, we assume:

Assumption 4 *Each component of the stochastic process $p(t)$ is first observed on a partition \mathcal{P}_{N_i} , $i = 1, 2$, then interpolated to the same evenly spaced partition \mathcal{P}_M according to the previous-tick interpolation (2.4).*

Remark that, after being interpolated on the partition \mathcal{P}_M , the time series $\{q_i(m)\}$ do not any longer include information on the original partitions \mathcal{P}_{N_i} . We study covariance estimation of $\{q_i(m)\}$ only.

Realized covariance on partition \mathcal{P}_M is defined by

$$RC(M) = \sum_{m=1}^M [q_1(m) - q_1(m-1)] [q_2(m) - q_2(m-1)], \quad (2.5)$$

and this is used as an estimator of $\int_0^T \Omega_{12}(t) dt$. However, when using interpolated data, $RC(M)$ is biased toward zero. This phenomenon is known as the Epps effect (Epps, 1979). We provide an explanation for it using a simple example in the following sections.

3 Nonsynchronous bias and lead-lag modification

To understand the source of the bias, consider the following simple example. In Figure 1 a realization of $p(t)$ is drawn¹. In this case, the whole period is divided into three equidistant periods. At the bottom of the figure, we shows the time position of the previous ticks for each equidistant period. Now define the interval I_m as

$$I_m \equiv [t_{m-1}^1, t_m^1] \cap [t_{m-1}^2, t_m^2]$$

where t_m^1, t_m^2 are defined in (2.3).

By the independence of increments of Brownian motion, the expectation of the realized covariance is calculated as

$$\begin{aligned} \mathbb{E}[RC(M)] &= \sum_{m=1}^M \mathbb{E}[(q_1(m) - q_1(m-1))(q_2(m) - q_2(m-1))] \\ &= \sum_{m=1}^M \mathbb{E}[(p_1(t_m^1) - p_1(t_{m-1}^1))(p_2(t_m^2) - p_2(t_{m-1}^2))] \\ &= \sum_{m=1}^M \mathbb{E}\left[\int_{t_{m-1}^1}^{t_m^1} (\sigma_{11}(t)dW_1(t) + \sigma_{12}(t)dW_2(t)) \int_{t_{m-1}^2}^{t_m^2} \sigma_{22}(t)dW_2(t)\right] \\ &= \sum_{m=1}^M \int_{I_m} \Omega_{12}(t) dt \end{aligned}$$

¹Remark that in our discussion we focus on *ex post* inference conditional on covariance path and arrival time P_{N_i} , so we can treat them as deterministic.

Figure 1: $M = 3$, $N_1 = 7$, $N_2 = 5$.

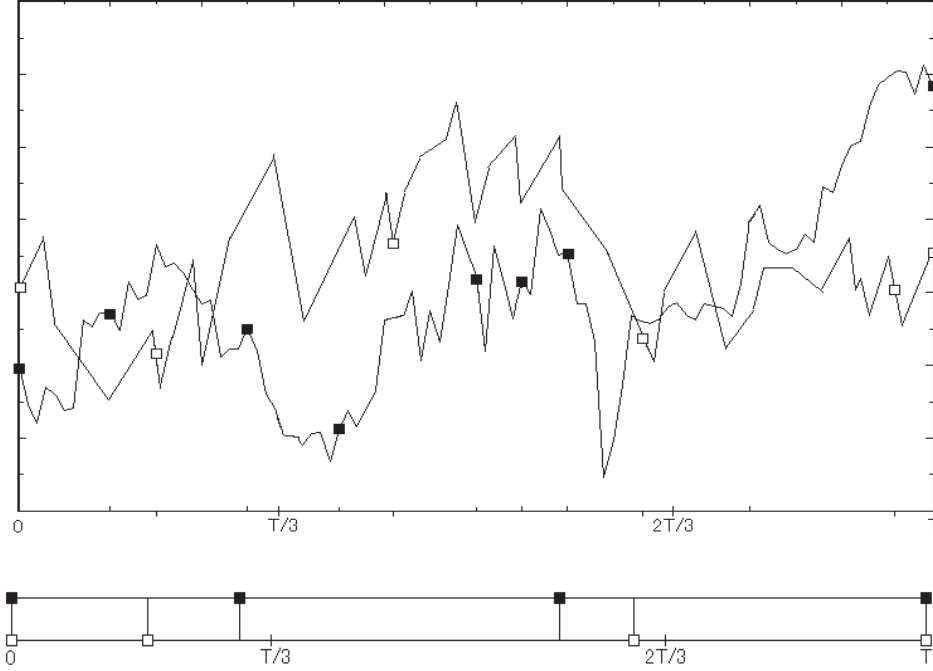


Figure 2: Intersection intervals



Thus, $RC(M)$ can account for the interval $\bigcup_{m=1}^M I_m$ only.² Figure 2 shows the intervals I_m in the case of Figure 1. The two gaps the expectation cannot cover is the source for the bias towards zero. We call this bias *nonsynchronous bias*.

To account for the *nonsynchronous bias*, in other words, to fill up the gaps, it is enough to use the following estimator.³

$$RCLL(L, U, M) = \sum_{m=1}^M \sum_{k=-L}^U [q_1(m+k) - q_1(m+k-1)] [q_2(m) - q_2(m-1)], \quad (3.1)$$

²If observations are synchronous, this interval coincides with $[0, T]$ making $RC(M)$ unbiased.

³This estimator has been proposed in the financial literature, see e.g. Scholes and Williams (1977); Lo and MacKinlay (1990), although they consider more specific models for estimating covariance.

where $q_1(m+k) - q_1(m+k-1) = 0$ if $m+k > M$ or $m+k-1 < 0$.⁴ The *RCLL* estimator simply adds lead and lag terms, accounting for the nonsynchronous bias.

In our example, *RCLL*(1, 1, 3) corrects the nonsynchronous bias. Define $\Delta q(m) = q(m) - q(m-1)$.

$$RCLL(1, 1, 3) = RC(3) + \sum_{m=2}^3 \Delta q_1(m) \Delta q_2(m-1) + \sum_{m=2}^3 \Delta q_1(m-1) \Delta q_2(m),$$

where $\Delta q_1(1) \Delta q_2(2)$ and $\Delta q_1(3) \Delta q_2(2)$ fill the first and second gap respectively. The products $\Delta q_1(2) \Delta q_2(1)$ and $\Delta q_1(2) \Delta q_2(3)$ are redundant and just increase the variance of the estimator. If we know the time of previous tick t_m^1 and t_m^2 , then we can choose one of the two $\Delta q_1(m) \Delta q_2(m-1)$ or $\Delta q_1(m-1) \Delta q_2(m)$ to cover m -th gap. If $t_m^1 < t_m^2$ ($t_m^1 > t_m^2$), we would add $\Delta q_1(m) \Delta q_2(m-1)$ ($\Delta q_1(m-1) \Delta q_2(m)$) only. However, when observation instants are not available, we are compelled to add all the cross terms.

4 Zero-return bias and bias-corrected estimator

In the above section, we have shown that *RCLL*(1, 1) is able to correct the nonsynchronous bias. However, also the *RCLL*(1, 1) estimator is still biased in the following case. For the same realization in the previous example we divide $[0, T]$ into six evenly spaced periods, as shown in Figure 3. By definition of previous tick interpolation, in this example we have $q_2(1) = q_2(0)$ and $q_2(5) = q_2(4)$, then $I_1 = \emptyset$ and $I_5 = \emptyset$. As shown at the bottom of Figure 3 uncovered gaps enlarge. In general, the area that is accounted for by *RC*(M) shrinks when increasing M not only because of the increasing number of gaps but also because of zero returns. Under high frequency situation, the existence of zero returns become more prominent. We call such bias *zero-return bias*.

Moreover in this case the modification by *RCLL*(1, 1) is not enough to cover whole interval. Figure 4 shows the area *RCLL*(1, 1, 6) covers. There is still a gap. Of course in this case *RCLL*(2, 2, 6) can cover whole interval, however, too many additional terms make the estimator noisy. In general, larger U and L makes the estimator less biased but more noisy. For an extreme example, *RCLL*(M, M, M) = $(p_1(T) - p_1(0))(p_2(T) - p_2(0))$ is always unbiased but very noisy. Instead of adding a fixed number of lead and lag terms at every m , we propose a flexible modification as follows.

Before the lead-lag type modification, we need another step of modification for zero return. For the case $M = 6$, we first consider a modification to account for the zero-return

⁴This condition is not necessary if we can arbitrary use the past data $q_i(-1), q_i(-2), \dots$ and the future data $q_i(M+1), q_i(M+2), \dots$

Figure 3: $M = 6, N_1 = 7, N_2 = 5$.

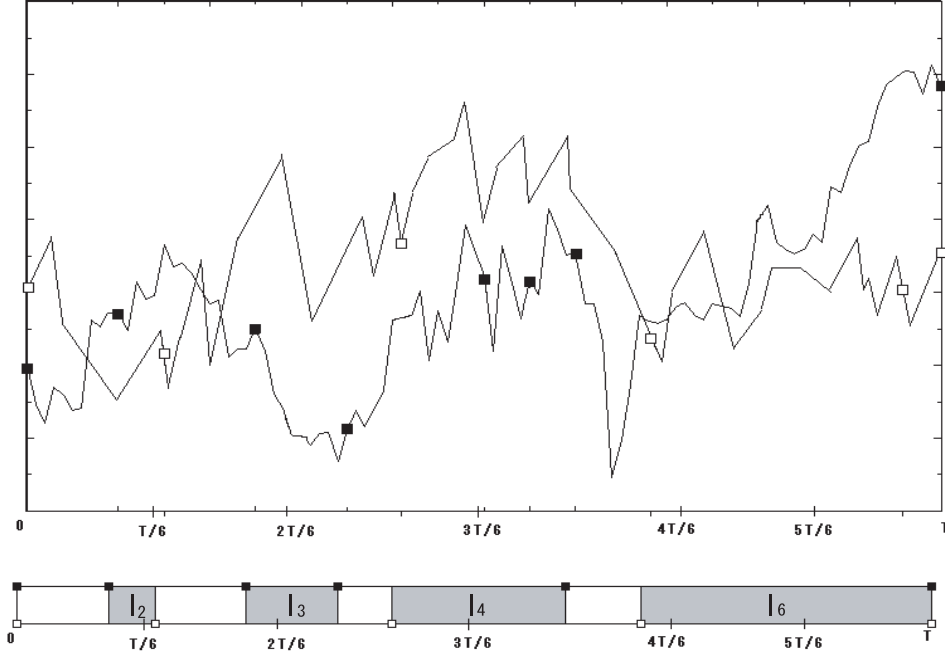
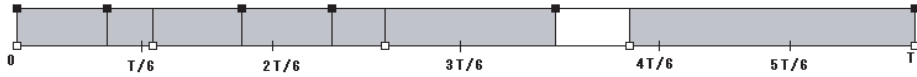


Figure 4: Area covered by $RCLL(1, 1, 6)$



bias. This can be done just by discarding zero returns. In other words we focus on price change vectors. Denote by $\Delta^z q(m) = q(m) - q(m - z)$. Consider price change vectors:

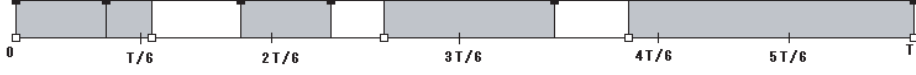
$$\begin{pmatrix} \Delta q_1(1), & \Delta q_1(2), & \Delta q_1(3), & \Delta q_1(4), & \Delta^2 q_1(6) \\ \Delta^2 q_2(2), & \Delta q_2(3), & \Delta q_2(4), & \Delta^2 q_2(6) \end{pmatrix},$$

then sum up cross products of two time-overlapping price changes:

$$\begin{aligned} & \Delta q_1(1)\Delta^2 q_2(2) + \Delta q_1(2)\Delta^2 q_2(2) + \Delta q_1(3)\Delta q_2(3) \\ & + \Delta q_1(4)\Delta q_2(4) + \Delta^2 q_1(6)\Delta^2 q_2(6). \end{aligned}$$

The expectation of this covers the area shown in figure 5. In order to fill up the three gaps in the figure, add the lead-lag terms for each gap

Figure 5: After modification for zero returns



$$\begin{array}{ccc} \Delta q_1(2)\Delta q_2(3) & \Delta q_1(3)\Delta q_2(4) & \Delta q_1(4)\Delta^2 q_2(6) \\ \Delta^2 q_2(2)\Delta q_1(3) & \Delta q_2(3)\Delta q_1(4) & \Delta q_2(4)\Delta^2 q_1(6) \end{array}$$

Now the whole interval is covered up, in other words, the modification is completed.

Similarly, in the general case, the modification consists of two steps: the modification for the zero-return bias and for the non-synchronous bias. With this in mind, we can easily derive the bias corrected estimator:

$$\begin{aligned} BC(M) &= \sum_{m_1=1}^M \sum_{m_2=1}^M 1_{A_1}[q_1(m_1) - q_1(m_1^-)] 1_{A_2}[q_2(m_2) - q_2(m_2^-)] 1_A \\ &\begin{cases} m_i^- = \max\{m < m_i : q_i(m) \neq q_i(m-1)\}, \quad i = 1, 2 \\ A_i = \{q_i(m_i) \neq q_i(m_i - 1)\}, \quad i = 1, 2 \\ A = \{[m_1^-, m_1] \cap [m_2^-, m_2] \neq \emptyset\} \end{cases} \end{aligned} \quad (4.1)$$

The term $1_{A_1}(q_1(m_1) - q_1(m_1^-))$ means that we include only nonzero-returns. The term 1_A takes cross product of time-overlapping pair of price changes as well as lead-lag one. By definition to cover the whole interval, the corrected estimator (4.1) is unbiased. Remark that $BC(M) = RCLL(1, 1, M)$ when there is no zero return in each interpolated return.

The bias-corrected estimator is similar in spirit to the unbiased covariance estimator proposed in Hayashi and Yoshida (2005), which is defined as:

$$\begin{aligned} HY &= \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} [p_1(t_{n_1}) - p_1(t_{n_1-1})] [p_2(t_{n_2}) - p_2(t_{n_2-1})] 1_B \\ B &= \{(t_{n_1-1}, t_{n_1}] \cap (t_{n_2-1}, t_{n_2}] \neq \emptyset\}. \end{aligned} \quad (4.2)$$

The HY estimator is designed for using all the observations, while $BC(M)$ is for the situation where data have been interpolated, thus we cannot obtain precise information of ticks including time stamps. However, these two estimators turn out to be the same for large M while RC and $RCLL$ shrink to almost zero.

Proposition 5 *For given completely asynchronous partitions $(\mathcal{P}_{N_1}, \mathcal{P}_{N_2})$, there exists $M^* \in \mathbb{N}$ such that $RC(M^*) = \Delta p_1(T)\Delta p_2(T)$ and $BC(M^*) = HY$.*

Furthermore, *RCLL* is designed for correcting bias, however, for larger M than M^* of the proposition above, *RCLL* provides the same estimates as *RC*.

Corollary 6 *For given completely asynchronous partitions $(\mathcal{P}_{N_1}, \mathcal{P}_{N_2})$, there exists $M^{**} \in \mathbb{N}$ such that $RCLL(L, U, M^{**}) = \Delta p_1(T) \Delta p_2(T)$.*

5 Monte Carlo study

In this section, we compute the *RC*, *RCLL*(1, 1), *RCLL*(2, 2), and *BC* on simulated time series. The model we simulate is:

$$\begin{pmatrix} dp_1(t) \\ dp_2(t) \end{pmatrix} = \begin{pmatrix} \sigma_{11}(t) & \sigma_{12}(t) \\ 0 & \sigma_{22}(t) \end{pmatrix} \begin{pmatrix} dW_1(t) \\ dW_2(t) \end{pmatrix}, \quad 0 \leq t \leq T \quad (5.1)$$

with

$$d\sigma_{ij}(t) = \kappa(\theta - \sigma_{ij}(t))dt + \gamma dW_{ij}(t), \quad i, j = 1, 2. \quad (5.2)$$

where $\kappa = 0.01$, $\theta = 0.01$, and $\gamma = 0.001$ for any i, j . We discretize the process (5.1-5.2) using a first-order Euler discretization scheme, with $\Delta t = 1$ second, for a total of $T = 60 \times 60 \times 4.5$ seconds. The data are observed with time differences which are drawn from an exponential distribution with mean $1/\lambda_1 = 1/0.04267 \approx 23.4$ seconds for p_1 and $1/\lambda_2 = 1/0.04787 \approx 20.9$ seconds for p_2 ,⁵ then interpolated to evenly spaced grids with different values M . We compare the performances of *BC*(M), *RCLL*(1, 1, M), *RCLL*(2, 2, M) and *RC*(M) for each M , as well as the *HY* estimator from original observations. We choose $M = 9, 18, 27, 54, 135, 270, 540, 1620, 3240$, and 16200 , which correspond to $T = 16200$, to 30, 15, 10, 5, 2, 1 minutes and 30, 10, 5, 1 seconds, respectively. We replicate 500 ‘daily’ experiments.

Since we know, on each trajectory, the integrated value of $\Omega_{12}(t)$, for each value of M we can draw the distribution histograms of the errors of estimators. Such distributions are shown in Figure 6, while table 1 reports the (simulated) sample MSE and average bias of the estimators. In Table 1, we also show the probability⁶

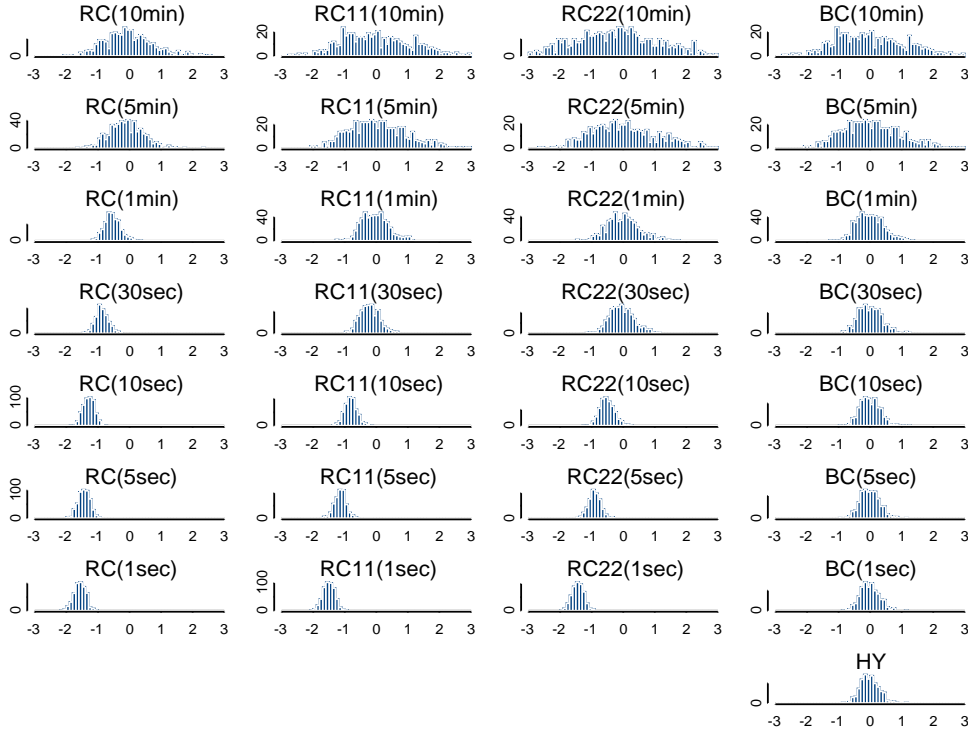
$$\begin{aligned} P_M &\equiv P(\{\Delta q_1(m) = 0\} \text{ or } \{\Delta q_2(m) = 0\}) \\ &= (1 - \lambda_1)^{T/M} + (1 - \lambda_2)^{T/M} - (1 - \lambda_1)^{T/M} (1 - \lambda_2)^{T/M}. \end{aligned} \quad (5.3)$$

It shows the expected percentage of the zero-return-bin in which either of two returns is zero.

⁵These values are typical on the intraday stock market

⁶In continuous time, the probability is $\exp(-\lambda_1 T/M) + \exp(-\lambda_2 T/M) - \exp(-\lambda_1 T/M) \exp(-\lambda_2 T/M)$. Since we discretized second by second, (5.3) is the exact probability in our simulation.

Figure 6: Distribution of errors in estimating covariances of $BC(M)$, $RC(M)$, $RCLL(1, 1, M)$, $RCLL(2, 2, M)$ and HY , for different values of M , see the text. The distributions are computed from 500 ‘daily’ replications.



Under our simulation design, the covariance between the first and the second asset is positive on average: $\Omega_{12}(t)$ reverts around a positive mean of 0.0001 since $\Omega_{12}(t) = \sigma_{11}(t)\sigma_{12}(t)$, and both $\sigma_{11}(t)$ and $\sigma_{12}(t)$ revert around a mean of 0.01. Thus we expect a downward-biased covariance estimate when using RC or $RCLL$.

Our simulations show, visually, the results obtained theoretically in the previous sections. The RC is biased (Figure 6), and the bias increases with increasing M . The same happens with $RCLL$, both with $L = U = 1$ and $L = U = 2$. Instead, BC is unbiased at all M , and its precision increases with increasing M , approaching the precision of the HY estimator for M large enough. In this case, BC and HY are not equal, as in Proposition 5 for large M , since in our sample synchronous observations happen with an intensity of $0.04267 \cdot 0.04787 \approx 0.2\%$ per second, that is we simulate on average ≈ 33 synchronous observations on each trajectory. By definitions of BC and $RCLL(1, 1)$, both estimators provide close estimates at lower frequencies, in which the number of zero returns is low.

In Table 1, the MSE of both estimators are the same up to 5 minutes ($M = 54$).

The advantage of using bias corrected estimators instead of RC depends on M and on the frequencies of observations N_1 and N_2 . At low frequencies, the increase of variance from additional terms overwhelms bias correction effects. In practice, bias corrected estimators should be used if RC shrinks to zero.⁷ Summarizing, at lower frequencies BC and $RCLL(1, 1)$ provide close results. At higher frequencies (or, when the number of zero returns is larger) the advantage of using BC becomes clear.

6 Covariance of high frequency data

We apply the BC estimator to compute the daily covariance from intraday data of individual stock prices (Honda, Nissan, and Toyota). The data are gathered from Japanese stock exchanges.⁸ We obtained 1-minute previous-tick-interpolated data (just taking closing prices of 1 minute bins) from July 2 to September 28, 2001 (63 trading days).⁹ To compare $BC(M)$ with $RC(M)$, $RCLL(1, 1, M)$, and $RCLL(2, 2, M)$, we compute averages of daily covariances for frequencies ranging from one minute to one hour (Figures 7,8 and 9). Such plots are called covariance (or volatility) signature plots in the financial econometrics literature. At the bottom of the figures, we show the ratio of zero-return bins: the ratio of the number of bins in which return of either of two assets is zero to total number of bins, showing that the actual data contain a significant amount of zero-returns not only because of no-trading bins but also because of price discreteness. RC computes smaller estimates than other estimators all over the intra-hourly frequencies. Difference between BC and $RCLL(1, 1)$ becomes clear at less than 10 (Honda-Nissan) or 15 (Honda-Toyota, Nissan-Toyota) minutes. As for $RCLL(2, 2)$, the difference also becomes clear at less than 5 (Honda-Nissan) or 10 (Honda-Toyota, Nissan-Toyota) minutes. The robustness of BC with respect to frequency supports our theoretical results.

We also analyzed the behavior of BC at time scales lower than one minute on high frequency data belonging to the TAQ database. We find that BC is biased toward zero in the very high frequency regime in a similar fashion with respect to the other estimators. This is not surprising, since also the HY estimator is biased in this situation, as shown

⁷It is important to remark that in our setting there is no microstructure noise, which is important for high frequency financial data, see Griffin and Oomen (2006). Microstructure noise has been considered in Bandi and Russell (2005); Zhang (2006); Palandri (2006). However, the independent noise they assume in their papers does not have any impact on the unbiasedness of our estimator.

⁸The trading hour of one day is 4.5 hours.

⁹The numbers of recorded transactions are 54,886, 49,298, and 66,544 for Honda, Nissan and Toyota, respectively. It means average duration times are 18.6, 20.7, and 15.3 seconds, respectively.

for example by Griffin and Oomen (2006). It is likely that our diffusion model (even with independent noise) cannot capture the dynamics of stock prices for extremely high frequencies.¹⁰

7 Concluding remarks

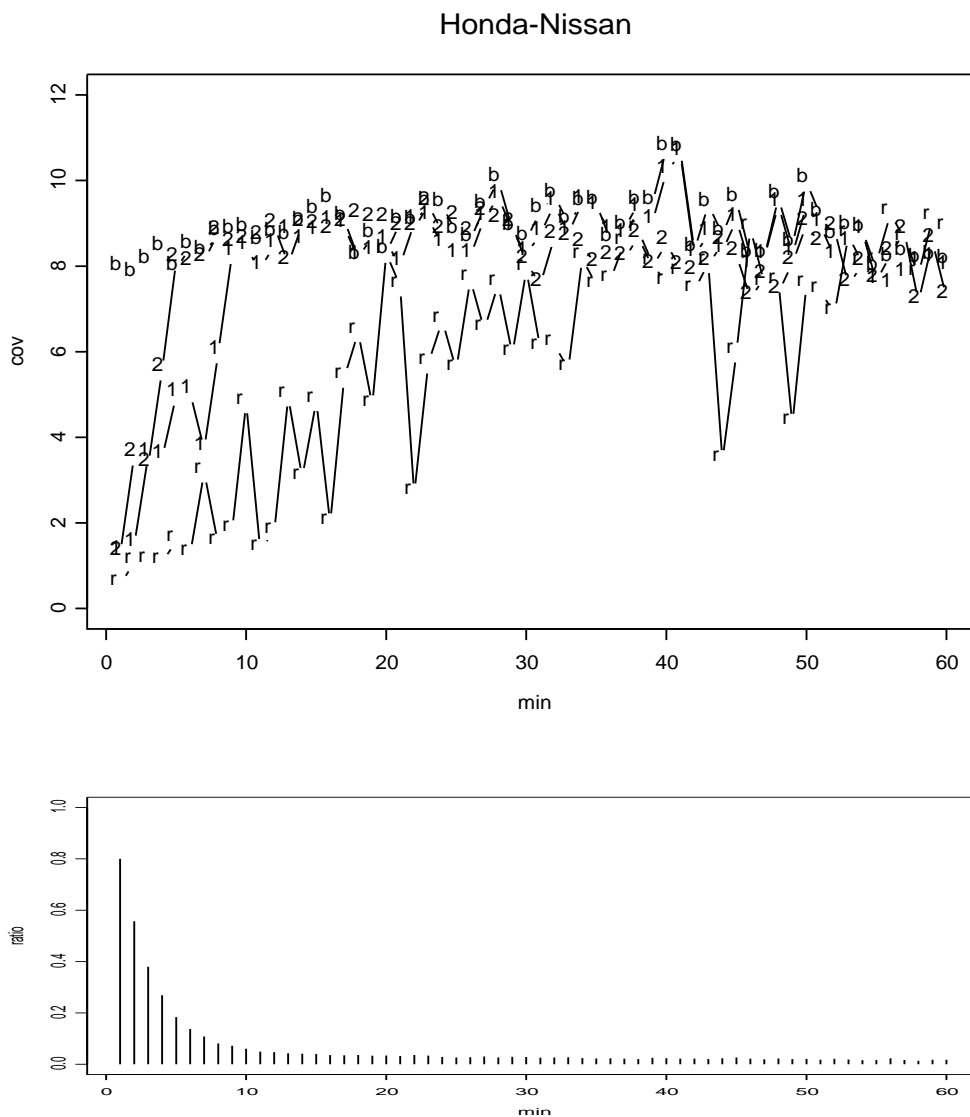
In this paper, we analyze the problem of covariance estimation when data are interpolated to an evenly spaced grid using the last available observation. We point out that using realized covariance leads to biased estimate, and that the bias depends on two sources: non-synchronous bias and zero-return bias. We also show that traditional lead-lag methods cannot account for zero-return bias. We then propose a bias corrected estimator which corrects for both sources. Our new estimator guarantees the unbiasedness at every frequency.

The proposed estimator can be helpful in applications especially under the situation where portfolio manager faces high frequency interpolated data of various kind of assets including some illiquid assets or periods. When time series include a significant amount of zero returns, our estimator should be implemented.

It is important to remark that in this paper we do not study the impact of microstructure noise since we focus on the unbiasedness of the estimator. Although the independent noise assumed in the related literatures does not have any impact on unbiasedness, it does have significant impact on efficiency. Clearly, the extension to this direction is crucial and it is now under development.

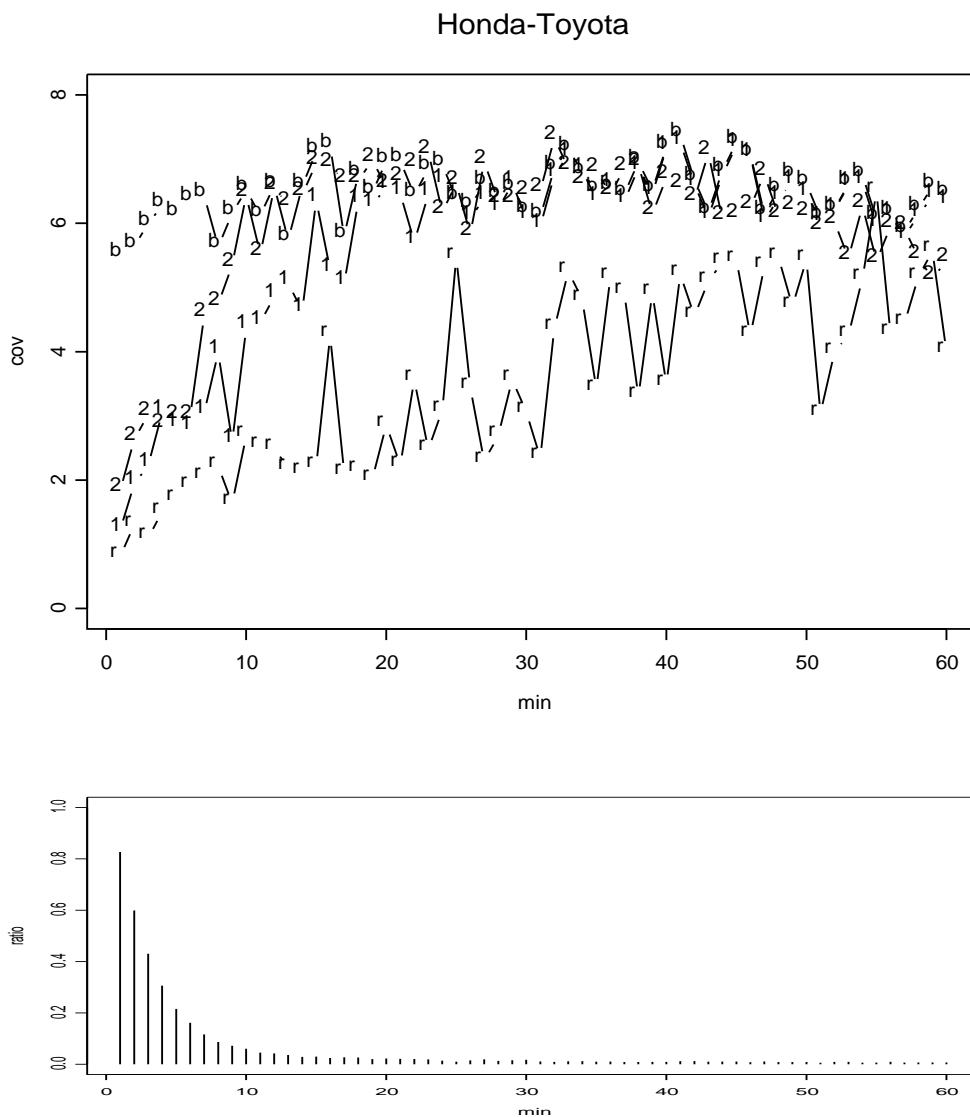
¹⁰Griffin and Oomen (2006) report that *HY* shrink to zero at less than 2 – 3 ticks for quotes data, 10 ticks for transaction data.

Figure 7: Daily covariance ($\times 10^4$) signature plots (upper) and zero-return ratio (bottom)



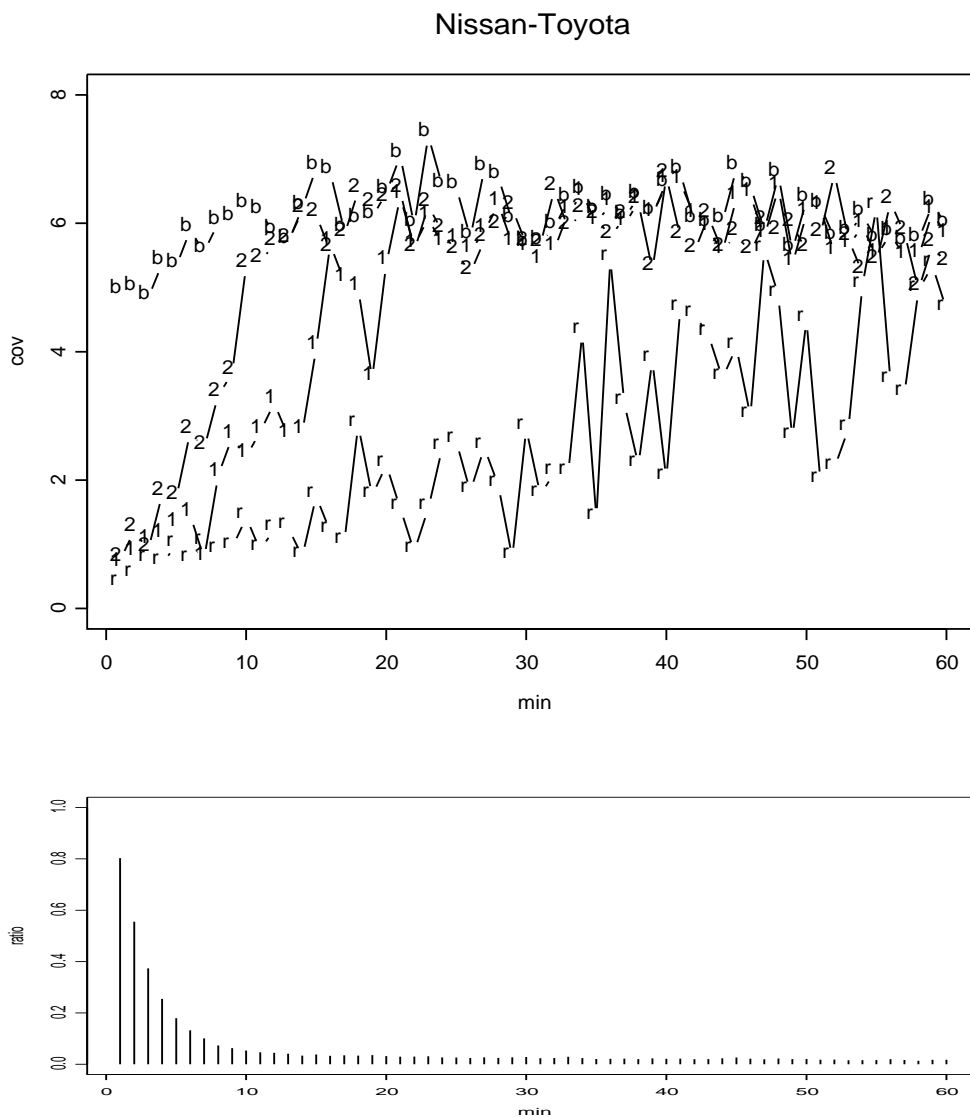
Note: “r” denotes 63 days mean of estimates by $RC(M)$, “1” by $RC(1,1, M)$, “2” by $RCLL(2,2, M)$, and “b” by $BC(M)$. The ratio is computed by (the number of bins in which either of two returns is zero)/(total number of bins).

Figure 8: Daily covariance ($\times 10^4$) signature plots (upper) and zero-return ratio (bottom)



Note: “r” denotes 63 days mean of estimates by $RC(M)$, “1” by $RC(1,1,M)$, “2” by $RCLL(2,2,M)$, and “b” by $BC(M)$. The ratio is computed by (the number of bins in which either of two returns is zero)/(total number of bins).

Figure 9: Daily covariance ($\times 10^4$) signature plots (upper) and zero-return ratio (bottom)



Note: “r” denotes 63 days mean of estimates by $RC(M)$, “1” by $RC(1,1, M)$, “2” by $RCLL(2,2, M)$, and “b” by $BC(M)$. The ratio is computed by (the number of bins in which either of two returns is zero)/(total number of bins).

A Proofs

Proof of Proposition 5. Since $\mathcal{P}_{N_1} \cap \mathcal{P}_{N_2} = \{0, T\}$, there exists a large M^* such that each bin $((m-1)T/M^*, mT/M^*]$ but $((M^*-1)T/M^*, T]$ includes at most one observation. In other words, for such M^* , every bin $((m-1)T/M^*, mT/M^*](m < M^*)$ must be one of the following three cases:

$$\text{only one observation of 1st asset } p_1(t_m^1), \quad (\text{A.1})$$

$$\text{only one observation of 2nd asset } p_2(t_m^2), \quad (\text{A.2})$$

$$\text{no observation.} \quad (\text{A.3})$$

For (A.1), there is no observation of 2nd asset, therefore, $q_2(m) = q_2(m-1)$. For (A.2), $q_1(m) = q_1(m-1)$. For (A.3), $q_1(m) = q_1(m-1)$ and $q_2(m) = q_2(m-1)$. Thus the $m(< M^*)$ th term of the $RC(M^*)$ must be zero for the every case. Now we obtain $RC(M^*) = \Delta q_1(M^*) \Delta q_2(M^*) = \Delta p_1(T) \Delta p_2(T)$.

For the same M^* , define the position of the bin that includes a observation as

$$\bar{m}_{n_i}^i \equiv \{m_i : \frac{(m_i-1)T}{M^*} < t_{n_i} \leq \frac{(m_i)T}{M^*}\}.$$

Then we have partitions

$$\mathcal{P}_{\bar{m}_1} \equiv \{0 = \bar{m}_0^1, \dots, \bar{m}_{n_1}^1, \dots, \bar{m}_{N_1}^1 = M^*\}$$

$$\mathcal{P}_{\bar{m}_2} \equiv \{0 = \bar{m}_0^2, \dots, \bar{m}_{n_2}^2, \dots, \bar{m}_{N_2}^2 = M^*\}$$

Notice that every element of $\mathcal{P}_{\bar{m}_1}$ and $\mathcal{P}_{\bar{m}_2}$ has one-to-one correspondence to that of \mathcal{P}_{N_1} and \mathcal{P}_{N_2} respectively. Now we can write $BC(M^*)$ using $\bar{m}_{n_i}^i$,

$$BC(M^*) = \sum_{n_1, n_2} [q_1(\bar{m}_{n_1}^1) - q_1(\bar{m}_{n_1-1}^1)][q_2(\bar{m}_{n_2}^1) - q_2(\bar{m}_{n_2-1}^1)] 1_{A'}$$

where $A' = \{[\bar{m}_{n_1-1}^1, \bar{m}_{n_1}^1] \cap [\bar{m}_{n_2-1}^1, \bar{m}_{n_2}^1] \neq \emptyset\}$. However, A' is equivalent with

$$A'' = \{[\frac{(\bar{m}_{n_1-1}^1 - \varepsilon_{n_1-1})T}{M^*}, \frac{(\bar{m}_{n_1}^1 - \varepsilon_{n_1})T}{M^*}] \cap [\frac{(\bar{m}_{n_2-1}^1 - \varepsilon_{n_2-1})T}{M^*}, \frac{(\bar{m}_{n_2}^1 - \varepsilon_{n_2})T}{M^*}] \neq \emptyset\}$$

for any $0 \leq \varepsilon_{n_i} < 1$. Since we can choose an ε_{n_i} such that

$$t_{n_i} = \frac{(\bar{m}_{n_i}^i - \varepsilon_{n_i})T}{M^*},$$

B is equivalent with A'' . By definition, $q_i(\bar{m}_{n_i}^i) = p_i(t_{n_i})$, now we obtain

$$BC(M^*) = \sum_{n_1, n_2} [p_1(t_{n_1}) - p_1(t_{n_1-1})][p_2(t_{n_2}) - p_2(t_{n_2-1})] 1_B.$$

■

Proof of Corollary 6. Define $K \equiv \max(L, U)$. We can choose a large M^{**} such that each bin has at most one observation and any bin of (A.1) or (A.2) is adjacent to at least K successive bins of (A.3). Then all terms of $RCLL(L, U, M^{**})$ but $[q_1(M^{**}) - q_1(M^{**} - 1)][q_2(M^{**}) - q_2(M^{**} - 1)]$ is zero. ■

References

- Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* 96(453).
- Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica* 71, 579–625.
- Bandi, F. M. and J. R. Russell (2005). Realized covariation, realized beta and microstructure noise. *mimeo*.
- Barndorff-Nielsen, O. E. and N. Shephard (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica* 72, 885–925.
- Cohen, K. J., G. A. Hawawini, S. F. Maier, A. Schwartz, and D. K. Whitcomb (1983). Friction in the trading process and the estimation of systematic risk. *Journal of Financial Economics* 12, 263–278.
- Epps, T. W. (1979). Comovements in Stock Prices in the Very Short Run. *Journal of the American Statistical Association* 74(366), 291–298.
- Griffin, J. E. and R. C. A. Oomen (2006). Covariance measurement in the presence of nonsynchronous trading and market microstructure noise. *mimeo*.
- Hayashi, T. and N. Yoshida (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli* 11, 359–379.
- Lo, A. W. and A. C. MacKinlay (1990). An econometric analysis of nonsynchronous trading. *Journal of Econometrics* 45, 181–211.
- Palandri, A. (2006). Consistent realized covariance for asynchronous observations contaminated by microstructure noise. *mimeo*.

Scholes, M. and J. Williams (1977). Estimating beta from nonsynchronous data. *Journal of Financial Economics* 5, 309–327.

Zhang, L. (2006). Estimating covariation: Epps effect, microstructure noise. *mimeo*.

Table 1: Sample MSE (bias) from 500 ‘daily’ replications

	M	RC	$RCLL(1,1)$	$RCLL(2,2)$	BC	P_M
30 min	9	1.77 (0.0769)	4.86 (0.0281)	7.21 (-0.0967)	4.86 (0.0281)	8.15e-035
15 min	18	0.893 (0.0384)	2.587 (0.0673)	4.183 (0.0540)	2.587 (0.0673)	9.09e-018
10 min	27	0.548 (-0.0272)	1.624 (0.109)	2.776 (0.0150)	1.624 (0.109)	4.5e-012
5 min	54	0.304 (-0.103)	0.845 (0.0633)	1.414 (0.103)	0.845 (0.0633)	2.49e-006
2 min	135	0.197 (-0.284)	0.327 (0.00941)	0.592 (0.0333)	0.328 (0.0112)	0.0081
1 min	270	0.354 (-0.540)	0.176 (-0.0343)	0.268 (-0.0000135)	0.180 (-0.00622)	0.122
30 sec	540	0.801 (-0.872)	0.135 (-0.214)	0.149 (-0.0477)	0.127 (0.0000881)	0.438
10 sec	1620	1.7031 (-1.293)	0.6968 (-0.811)	0.3162 (-0.512)	0.0926 (-0.00684)	0.863
5 sec	3240	2.1139 (-1.442)	1.3468 (-1.146)	0.8551 (-0.906)	0.0893 (-0.0059)	0.957
1 sec	16200	2.5209 (-1.577)	2.3059 (-1.507)	2.1103 (-1.441)	0.0867 (-0.00977)	0.998
HY					0.0833 (-0.011)	