Unbiased covariance estimation with interpolated data

Taro Kanatani†‡

JSPS Research Fellow, Institute of Economic Research, Kyoto University

Yoshida Honmachi, Sakyo-Ku, Kyoto 6068501, Japan

October 14, 2006

Abstract

In this paper we focus on the realized covariance bias when we are compelled to use evenly spaced data which have already been manipulated by previous-tick interpolation. We explain the cause of the bias and propose a bias-correction estimator. The proposed estimator is

---

*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” and 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

‡URL: http://www.geocities.jp/kanatanit/ (The latest version is available here.)
proved to be consistent with the Hayashi and Yoshida (2005)’s estimator under extremely high frequency situation. Numerical examples of both simulation and empirical study are also provided.

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

**JEL Classification:** C14; C32; C63

1 **Introduction**

In this paper we study a methodology for measuring covariance between financial assets. The covariance is widely called co-volatility or cross-volatility in the financial econometrics literature. High frequency financial data have been made easily available by recent developments of computer technology. Under such circumstances, it is natural to measure the volatility matrix by the sum of outer products of returns. The sum is referred to as realized volatility matrix, which has been theoretically and empirically investigated by many researchers. See e.g. Andersen, Bollerslev, Diebold, and Labys (2001), Andersen, Bollerslev, Diebold, and Labys (2003), Barndorff-Nielsen and Shephard (2004).

We focus on the off-diagonal elements of the realized volatility matrix, which is called realized covariance or realized covariation in this field. Since the empirical study by Epps (1979), the bias of realized covariance is known as Epps effect. The realized covariance is biased toward zero as the sampling frequency increases. This paper explains the cause of this phenomena under a simple settlings and propose a practical method for correcting the bias.
The rest of the paper is organized as follows. In the following subsection, we introduce the data generating process we assume throughout the paper and take a look at the realized covariance matrix. In Section 2, we study the bias of realized covariance and propose a bias-corrected estimator. In Section 3, we confirm our theory through a Monte Carlo study. We also present an empirical example in Section 4. The final section notes a conclusion and remaining works.

1.1 Data generating process and realized covariance matrix

We consider $n$-dimensional vector of logarithmic asset price $p(t)$ for $t \geq 0$. We assume that $p$ is a continuous stochastic volatility semimartingale (SVSM) with zero drift.\footnote{See Barndorff-Nielsen and Shephard (2004) for the SVSM.}

$$p(t) = \int_0^t \Sigma(u)dz(u),$$

where $\Sigma$ has elements that are all cadlag and $z$ is a vector standard Brownian motion. We set the drift vector as 0, for the purpose of simplification.\footnote{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.} We define the instantaneous or spot covariance matrix as

$$\Omega(t) \equiv \Sigma(t)\Sigma(t)',$$
that is to say, cross volatility between \( i \)th and \( j \)th asset is denoted as the \((i, j)\)th element of \( \Omega \):

\[
\omega_{ij}(t) = \sum_{n'=1}^{n} \sigma_{in'}(t) \sigma_{jn'}(t).
\]

Our target is not spot covariance matrix but integrated (cross) volatility (matrix) over \([0, T]\):

\[
\int_{0}^{T} \omega_{ij}(t) dt.
\]

For estimation of the integrated volatility, the following quadratic variation formula is a theoretical basis for using sum of outer product of return vector. If all assets are synchronously observed at the same time points

\[
0 = t_0 < t_1 < \cdots < t_N <= T,
\]

and

\[
\lim_{N \to \infty} \max_n (t_n - t_{n-1}) = 0
\]

then

\[
p - \lim_{M \to \infty} \sum_{n=1}^{N} (p(t_n) - p(t_{n-1}))(p(t_n) - p(t_{n-1}))' = \int_{0}^{T} \Omega(t) dt.
\]

See e. g. Barndorff-Nielsen and Shephard (2004).

However, each \( i \)th asset price is nonsynchronously traded or observed at different time points

\[
0 = t_{0i} < t_{1i} < \cdots < t_{Ni} <= T.
\]

We study the effect of such nonsynchronous trading (observation) on the realized covariance.
Since we concentrate on measuring the ex post cross volatility from a given observation and do not make any hypothesis on the structure of the underlying probability space, we can consider Σ(t) and tn as deterministic functions.

2 Realized covariance and bias corrected estimator

In this section we study the bias of the realized covariance and propose an unbiased estimator of the integrated covariance.

2.1 Previous-tick interpolation bias of realized covariance

Typically, raw data, which are unevenly spaced, are converted to evenly spaced data in order to apply to the usual discrete time series analysis. Dacorogna, Gençay, Müller, Olsen, and Pictet (2001) introduces some interpolation methods including previous tick interpolation. The previous-tick interpolation at time mT/M is defined by the following formula.

\[ q_i \left( \frac{mT}{M} \right) = p_i \left( \max \left\{ t_{n_i} : t_{n_i} \leq \frac{mT}{M} \right\} \right) \]  (2.1)

where max A and min A denote maximum and minimum elements of A, respectively.
Realized covariance is defined by

\[ RC_{12}(M) = \sum_{m=1}^{M} \left( q_1 \left( \frac{mT}{M} \right) - q_1 \left( \frac{(m-1)T}{M} \right) \right) \left( q_2 \left( \frac{mT}{M} \right) - q_2 \left( \frac{(m-1)T}{M} \right) \right) , \]

(2.2)
as an estimator of \( \int_0^T \omega_{12}(t) dt \). Without loss of generality, we limit our discussion to the case of two dimensional asset prices. For the purpose of simplification, we drop the suffix 12 and denote \( q_i(mT/M) \) as \( q_i(m) \). It is well known that the RC has the bias toward 0. The bias becomes more serious as \( M \) increases.

To see the cause of the bias, consider the following simple example. Consider the realization of two different log asset prices drawn in the figure 1. The bottom of the figure describes the time positions of the previous ticks. In this case, the whole period is divided into three equidistant periods. The figure shows the time position of previous tick of each period. This time position line enables us to easily understand the structure of the bias. Define the interval \( I_m \) as

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

where

\[ t_{im} = \max \left\{ t_n : t_n \leq \frac{mT}{M} \right\} . \]

Then notice that we can rewrite (2.1) as

\[ q_i(m) = p_i(t_{im}). \]

Figure 2 shows the intersection intervals of the example. By the inde
Figure 1: Nonsynchronous observations

Note: $M = 3$, $N_1 = 8$, $N_2 = 5$.

dependence of increment of Brownian motion, the expectation of the realized
covariance is calculated as

\[
E(\text{RC}(M)) = \sum_{m=1}^{M} E(q_1(m) - q_1(m-1))(q_2(m) - q_2(m-1))
\]

\[
= \sum_{m=1}^{M} E(p_1(t_1^m) - p_1(t_1^{m-1}))(p_2(t_2^m) - p_2(t_2^{m-1}))
\]

\[
= \sum_{m=1}^{M} E \left( \int_{t_1^{m-1}}^{t_1^m} \sum_{n'=1}^{n} \sigma_1(t)dz_{t}^{n'} \int_{t_2^{m-1}}^{t_2^m} \sum_{n'=1}^{n} \sigma_2(t)dz_{t}^{n'} \right)
\]

\[
= \sum_{m=1}^{M} \int_{I_m} \omega_{12}(t)dt
\]

where \(z_t^{n'}\) denotes \(n'\)th element of \(n\) dimensional Brownian motion \(z_t\). Thus \(\text{RC}\) can cover the interval of \(\bigcup_{m=1}^{M} I_m\). If observations are synchronous, such gaps never happen. We call this bias \textit{nonsynchronous bias}.

We then divide \([0, T]\) into 6 equidistant periods. By definition of previous tick interpolation, in this example, \(q_2(1) = q_2(0)\) and \(q_2(5) = q_2(4)\), then \(I_1 = \emptyset\) and \(I_5 = \emptyset\). Thus the area the RC accounts for shrinks by increasing \(M\). The gaps become larger when there is no observation in the bin, in other words, when either return is zero. We call such bias \textit{zero-return bias}. 
Figure 3: More serious case

Note: $M = 6$, $N_1 = 8$, $N_2 = 5$.

2.2 Bias corrected estimator

As shown in the example, the bias toward zero comes from the interval gaps RC cannot cover. When we have at least one pair of both asset data as seen in the example of $M = 3$, we can cover the gaps by add lead and lag terms. In the example, we can modify the RC as

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

$\Delta q_1(1)\Delta q_2(2)$ and $\Delta q_1(3)\Delta q_2(2)$ work to fill the first and second gap respectively. $\Delta q_1(2)\Delta q_2(1)$ and $\Delta q_1(2)\Delta q_2(3)$ are redundant and just increase the
variance of the estimator. However, unless we have time stamp of each tick, in other words, we cannot choose $\Delta q_1(m)\Delta q_2(m - 1)$ or $\Delta q_1(m - 1)\Delta q_2(m)$. In this paper we assume that we have no time stamp but have evenly spaced data which have already been manipulated by the previous tick interpolation.

For the case $M = 6$, we need the additional modification before the lead-lag modification. Since zero returns cause large gaps, we ignore zero returns, in other words, we consider price changes. In the example, consider

\[
\begin{array}{cccccccc}
\Delta q_1(1) & \Delta q_1(2) & \Delta q_1(3) & \Delta q_1(4) & \Delta q_1(5) & \Delta q_1(6) \\
\Delta^2 q_2(2) & \Delta q_2(3) & \Delta q_2(4) & \Delta^2 q_2(6)
\end{array}
\]

then sum cross products of two time-overlapped returns

\[
\Delta q_1(1)\Delta^2 q_2(2) + \Delta q_1(2)\Delta^2 q_2(2) + \Delta q_1(3)\Delta q_2(3)
\]

\[
+ \Delta q_1(4)\Delta q_2(4) + \Delta q_1(5)\Delta^2 q_2(6) + \Delta q_1(6)\Delta^2 q_2(6).
\]

The expectation of this cover the area shown in figure 4. The modification

Figure 4: After modification for zero returns

is completed by adding lead-lag terms

\[
\begin{array}{cccccccc}
\Delta q_1(2)\Delta q_2(3) & \Delta q_1(3)\Delta q_2(4) & \Delta q_1(4)\Delta^2 q_2(6) \\
\Delta^2 q_2(2)\Delta q_1(3) & \Delta q_2(3)\Delta q_1(4) & \Delta q_2(4)\Delta q_1(5)
\end{array}
\]

Similarly, in general case, the modification consists of two steps: (1) zero-return modification (2) lead-lag modification. We define a bias corrected
estimator by general form

\[ BC(M) = \sum_{m_1,m_2} 1_{A_1}(q_1(m_1) - q_1(m^-_1))1_{A_2}(q_2(m_2) - q_2(m^-_2))1_A \] (2.3)

where

\[ m^-_i = \max\{m'_i < m_i : q_i(m'_i) \neq q_i(m'_i - 1)\}, \] (2.4)

\[ A_i = \{q_i(m_i) \neq q_i(m_i - 1)\} \] (2.5)

\[ A = \{ [m^-_1, m_1] \cap [m^-_2, m_2] \neq \emptyset \} \] (2.6)

1_{A_i}(q_i(m_i) - q_i(m^-_i)) means that we pick up price changes. By 1_A we can sum up time-overlapping and lead-lag price changes. Note that the lead-lag products are taken into consideration by the end points. This is in the same line with the estimators for proposed by Hayashi and Yoshida (2005).

\[ HY = \sum_{n_1,n_2} (p_1(t_{n_1}) - p_1(t_{n_1-1}))(p_2(t_{n_2}) - p_2(t_{n_2-1}))1_B \]

where \( B = \{(t_{n_1-1}, t_{n_1}] \cap (t_{n_2-1}, t_{n_2}] \neq \emptyset \} \). Although \( HY \) is designed for transaction data, \( BC(M) \) is consistent with \( HY \) for a large \( M \).

**Theorem 1** If \( t_{n_1} \neq t_{n_2} \) for any \((n_1, n_2)\), then there exists a large \( M \) such that

\[ RC(M) = 0, \] (2.7)

\[ BC(M) = HY. \] (2.8)

**Proof.** See Appendix A ⚫
3 Monte Carlo study

We examine the above theory through a Monte Carlo study. Without loss of
generality, we set the number of assets as two. We follow the Monte Carlo
design of Barucci and Renò (2002) with some modification for multivariate
setting: we generate proxy for continuous observations by the Euler approx-
imation of the following stochastic differential equations with a time step of
one second,

\[
\begin{pmatrix}
dp_1(t) \\
dp_2(t)
\end{pmatrix} =
\begin{pmatrix}
\sigma_{11}(t) & 0 \\
\sigma_{21}(t) & \sigma_{22}(t)
\end{pmatrix}
\begin{pmatrix}
dz_1(t) \\
dz_2(t)
\end{pmatrix}, \quad 0 \leq t \leq T
\]

\[d\sigma_{ij}(t) = \kappa_{ij} (\theta_{ij} - \sigma_{ij}(t)) dt + \gamma_{ij} dW_{ij}(t), \quad i, j = 1, 2.\]

where \(\kappa_{ij} = 0.01, \theta_{ij} = 0.01, \) and \(\gamma_{ij} = 0.001\) for any \(i, j\) and \(T = 60 \times 60 \times 4.5\)
seconds. Time differences are drawn from an exponential distribution with
mean \(1/0.04267 \approx 23.4\) seconds for \(p_1\) and \(1/0.04787 \approx 20.9\) seconds for \(p_2.\)\(^3\)

\[F(t_k^i - t_{k-1}^i) = 1 - \exp\{ -\lambda_i (t_k^i - t_{k-1}^i) \}, \quad i = 1, 2\]

where \(F(\cdot)\) denotes a cumulative distribution function, \(\lambda_1 = 0.04267\) and
\(\lambda_2 = 0.04787.\) These values of intensity correspond with those of individual
stocks in the later empirical example.

We compared the performances of \(BC(M)\) and \(RC(M)\) for different \(M\)s.
Since \(T = 16200, \) \(M = 9, 18, 27, 54, 135, 270, 540, 1620, 3240, 16200\) corre-
spond 30 minutes, 15 minutes, 10 minutes, 5 minutes, 2 minutes, 1 minute,
30 seconds, 10 seconds, 5 seconds, and 1 second, respectively. We performed
500 ‘daily’ replications.

\(^3\)Of course, our method allows the duration to be correlated or autocorrelated. See
Engle and Russell (1998) for an autocorrelated duration model.
Figure 5 shows the distribution of errors of $RC(M)$ and $BC(M)$:

$$RC(M) - \int_{0}^{T} \omega_{12}(t)dt, \text{ and, } BC(M) - \int_{0}^{T} \omega_{12}(t)dt,$$

respectively.

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

<table>
<thead>
<tr>
<th></th>
<th>$RC(M)$</th>
<th>$BC(M)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>MSE</td>
<td>Bias</td>
</tr>
<tr>
<td>30 min</td>
<td>1.5430</td>
<td>-0.0878</td>
</tr>
<tr>
<td>15 min</td>
<td>0.8276</td>
<td>-0.0801</td>
</tr>
<tr>
<td>10 min</td>
<td>0.5742</td>
<td>-0.0744</td>
</tr>
<tr>
<td>5 min</td>
<td>0.2663</td>
<td>-0.1272</td>
</tr>
<tr>
<td>2 min</td>
<td>0.1832</td>
<td>-0.2750</td>
</tr>
<tr>
<td>1 min</td>
<td>0.3443</td>
<td>-0.5354</td>
</tr>
<tr>
<td>30 sec</td>
<td>0.7850</td>
<td>-0.8652</td>
</tr>
<tr>
<td>10 sec</td>
<td>1.6787</td>
<td>-1.2833</td>
</tr>
<tr>
<td>5 sec</td>
<td>2.0962</td>
<td>-1.4360</td>
</tr>
<tr>
<td>1 sec</td>
<td>2.5154</td>
<td>-1.5750</td>
</tr>
</tbody>
</table>

HY

<table>
<thead>
<tr>
<th></th>
<th>MSE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.08356</td>
<td>0.00617</td>
</tr>
</tbody>
</table>

Table 1 reports the sample MSE and bias of each estimator from 500
replications:

\[
\frac{1}{R} \sum_{r=1}^{R} \left( \text{Estimate}^{(r)} - \int_0^T \omega_{ij}^{(r)} (t) \, dt \right)^2
\]

and

\[
\frac{1}{R} \sum_{r=1}^{R} \left( \text{Estimate}^{(r)} - \int_0^T \omega_{ij}^{(r)} (t) \, dt \right)
\]

where \( r \) denotes the number of replications and \( R = 500 \). Under our simulation design, the correlation between the 1st and 2nd asset is on average positive: \( \omega_{12} (t) \) varies around a positive mean of 0.0001 because

\[
\omega_{12} (t) = \sigma_{11} (t) \sigma_{21} (t)
\]

and both of \( \sigma_{11} \) and \( \sigma_{21} \) have the mean of 0.01. Therefore, in this case, the bias toward zero means the downward bias. As expected from the previous discussion, the downward bias is getting more serious as the number of bins increase. We also calculated Hayashi and Yoshida (2005)’s estimator from raw data in order to compare \( RC(16200) \). Both estimates are not exactly the same. Since the time is not continuous, there exit synchronous observations at the probability 0.04267 \times 0.04787 per second.

4 An application for individual stock prices

We applied the estimators to high frequency data of Japanese individual stock prices (Honda and Nissan). We obtained 1-minute previous-tick-interpolated data from July 2 to September 10, 2001 (50 trading days). We calculated \( RC(M) \) and \( BC(M) \) at \( M = 9, 18, 27, 54, 135, 270 \). Since \( T = 16200 \), \( M = 9, 18, 27, 54, 135, 270 \) correspond 30 minutes, 15 minutes, 10 minutes, 5 minutes, 2 minutes, 1 minute, respectively.
Figure 6 plots daily estimates for integrated covariance by $RC(M)$ and $BC(M)$. For the easiness of presentation, we just draw results of $M = 9, 27, 135$ for the first 20 days. It is clear that $RC(M)$ shrinks to zero as $M$ increases.

To show this phenomenon more generally, we computed cumulative estimates of $RC(M)$ and $BC(M)$ until $i$th day:

$$\sum_{j=1}^{i} Estimate^{(j)}$$

where $Estimate^{(j)}$ denotes daily estimate of $RC(M)$ or $BC(M)$ of $j$th day. As shown in Figure 7, two stock prices are totally positively correlated, $RC(M)$ is seriously downward biased, and $BC(M)$ is robust to $M$.

### 5 Concluding remarks

In this paper we explain the cause of the bias of realized covariance which is calculated from interpolated data through previous tick interpolation. We propose a simple and practical method for correcting the bias, which is in the same line with unbiased estimator of Hayashi and Yoshida (2005). Both Monte Carlo study and empirical application support our theory. Since we concentrate on proposing an unbiasedness of the estimator, there is no variance analysis. Also, we ignore observation error (microstructure noise), because it does not have any impact to unbiasedness if the noise is uncorrelated between different assets. By incorporating these points, we can proceed to MSE analysis in the future study.
A  Proof of Theorem 1

Since $t_{n_1} \neq t_{n_2}$ for any $(n_1, n_2)$, there exists a large $M$ such that each bin $(m - 1, m]$ has at most one transaction. In other words, the bin $(m - 1, m]$ must include either

only one 1st asset transaction $p_1(t_m^1)$,  \hspace{1cm} (A.1)
only one 2nd asset transaction $p_2(t_m^2)$, or  \hspace{1cm} (A.2)
no transaction.  \hspace{1cm} (A.3)

For (A.1), there is no transaction of 2nd asset, in other words, $q_2(m) = q_2(m - 1)$. 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 $m$th term of the RC$(M)$ must be zero for the every case. Now we obtain (2.7).

For the same $M$, all time points $\{t_{n_i}\}$ is completely divided into $2M$ bins. Then $\{p_i(t_{n_i}) - p_i(t_{n_i-1})\}_{n_i=1}^{N_i}$ is exactly consistent with

$$\{q_i(m_i) - q_i(m^-_i), m_i = 1, \ldots, M : 1_{A_i} = 1\}.$$  

Thus $A$ and $B$ are completely equivalent since $(t_{n_1}, t_{n_2}) \in R_+^2$ is well expressed by discrete time points $0 \leq m_i \leq M, i = 1, 2$.

References


Figure 5: Distribution of errors

Note: RC15min: \( RC(18); \) RC10min: \( RC(27); \) RC5min: \( RC(54); \) RC2min: \( RC(135); \) RC1min: \( RC(270); \) RC30sec: \( RC(540); \) RC1sec: \( RC(16200); \) BC15min: \( BC(18); \) BC10min: \( BC(27); \) BC5min: \( BC(54); \) BC2min: \( BC(135); \) BC1min: \( BC(270); \) BC30sec: \( BC(540); \) BC1sec: \( BC(16200); \) HY: \( HY; \) The distribution is computed with 500 ‘daily’ replications.
Figure 6: Daily realized covariance

Note: Realized covariances (upper) and bias corrected estimators (lower) between Honda and Nissan. The figure is drawn for the first 20 days.
Figure 7: Cumulative realized covariance

Note: Cumulative realized covariances (left) and bias corrected estimators (right) between Honda and Nissan. The figure is drawn for the 50 business days from July 2 to September 10, 2001.