Unbiased Covariance Estimation with Interpolated Data

Taro Kanatani

JSPS Fellow, Institute of Economic Research, Kyoto University

January 26, 2007, Rimini
Introduction

- Estimation of covariance between financial assets (co-volatility, cross-volatility)
  ⇒ Portfolio risk, etc.

- High-frequency data (Intraday data)
e.g. Hourly data, 30 minutes data, ⋯, 5 minutes data
  ⇒ Realized Volatility, Realized Covariance

- However, RC has a serious problem:

  Bias toward 0 (Epps effect)
Epps effects

Correlation of Log Returns for Three Stocks

(Epps, 1979, JASA)

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Chrysler-Ford</th>
<th>Chrysler-GM</th>
<th>Ford-GM</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 minutes</td>
<td>-0.014</td>
<td>0.007</td>
<td>0.055</td>
</tr>
<tr>
<td>20 minutes</td>
<td>0.017</td>
<td>0.026</td>
<td>0.118</td>
</tr>
<tr>
<td>40 minutes</td>
<td>0.041</td>
<td>0.040</td>
<td>0.197</td>
</tr>
<tr>
<td>1 hour</td>
<td>0.023</td>
<td>0.065</td>
<td>0.294</td>
</tr>
<tr>
<td>2 hours</td>
<td>0.112</td>
<td>0.129</td>
<td>0.383</td>
</tr>
<tr>
<td>3 hours</td>
<td>0.361</td>
<td>0.518</td>
<td>0.519</td>
</tr>
</tbody>
</table>

‘The higher frequency, the less correlation’
Data Generating Process:

\[
\frac{dp(t)}{n_{\times 1}} = \sum(t) \cdot \frac{dz(t)}{n_{\times n} n_{\times 1}}, \quad 0 \leq t \leq T
\]

where \( z(t) \) is standard Brownian motion.

(Instantaneous or spot) volatility matrix:

\[
\Omega(t) \equiv \sum(t) \sum(t)'
\]

\( \omega_{ii}(t) \): volatility, \( \omega_{ij}(t) \): covolatility.
Estimation of integrated volatility $\int_0^T \Omega(t)dt$

$\Rightarrow$ Quadratic variation:

$$p - \lim_{M \to \infty} \sum_{m=1}^M \Delta p(mT/M) \Delta p(mT/M)' = \int_0^T \Omega(t)dt$$

Realized volatility matrix

Realized volatility diagonal
Realized covariance off-diagonal

Intuitively,

$$\int_0^T dp(t)dp(t)' = \int_0^T \Sigma(t) \underbrace{dz(t)dz(t)'}_{dtI_n} \Sigma(t)'$$

* If all assets are observed at the same time points $\{mT/M\}_{m=0}^M$ (synchronous and equidistant sampling), there is no problem...
Nonsynchronous trading (observation)

Financial assets are traded (observed) at different time points.

1st asset: \(0 = t_{01} < t_{11} < \cdots < t_{n_1} < \cdots < t_{N_1-1} < t_{N_1} = T\)

2nd asset: \(0 = t_{02} < t_{12} < \cdots < t_{n_2} < \cdots < t_{N_2-1} < t_{N_2} = T\)

Transaction data:

\[
\begin{align*}
N_{1+1} & \text{ observations} \\
p_1(t_{01}), p_1(t_{11}), \cdots, p_1(t_{n_1}), \cdots, p_1(t_{N_1-1}), p_1(t_{N_1}) \\
N_{2+1} & \text{ observations} \\
p_2(t_{02}), p_2(t_{12}), \cdots, p_2(t_{n_2}), \cdots, p_2(t_{N_2-1}), p_2(t_{N_2})
\end{align*}
\]
Nonsynchronous trading

■ 1st asset \((N_1 = 7)\); □ 2nd asset \((N_2 = 5)\)
Unevenly spaced data

\[ \{ p_i(t_{n_i}) \}_{n_i=0}^{N_i} \rightarrow \text{Interpolation} \rightarrow \{ q_i(mT/M) \}_{m=0}^{M} \]

Evenly spaced data

Previous-tick interpolation

\[ q_i \left( \frac{mT}{M} \right) = p_i \left( \max \left( t_{n_i} : t_{n_i} \leq \frac{mT}{M} \right) \right) \]
Previous-tick interpolation ($M = 3$)

- **Raw data;**
- **Interpolated data**

- ■ Raw data; □ Interpolated data
Realized covariance from the interpolated data

Realized covariance (An estimator of $\int_0^T \omega_{ij}(t) dt$)

$$RC(M) = \sum_{m=1}^{M} \left\{ q_i \left( \frac{mT}{M} \right) - q_i \left( \frac{(m-1)T}{M} \right) \right\} \left\{ q_j \left( \frac{mT}{M} \right) - q_j \left( \frac{(m-1)T}{M} \right) \right\}$$

⇒ The bias toward 0

$$E(RC(M)) \to 0 \text{ as } M \to \infty$$

⇒ Epps effect (Epps, 1979)
Time positions of previous ticks
For simple notation,

\[ q_i(m) \equiv q_i(mT/M), \Delta q_i(m) \equiv q_i(mT/M) - q_i((m - 1)T/M) \]

The expectation of realized covariance:

\[
E(RC(3)) = E\{\Delta q_1(1)\Delta q_2(1) + \Delta q_1(2)\Delta q_2(2) + \Delta q_1(3)\Delta q_2(3)\} = \int_{I_1} \omega_{12}(t)dt + \int_{I_2} \omega_{12}(t)dt + \int_{I_3} \omega_{12}(t)dt
\]
More serious case ($M = 6$)
\[ M = 6 \]

\[ q_1(4) = q_1(5), \quad q_2(0) = q_2(1), \quad q_2(4) = q_2(5) \]

\[ \Rightarrow \quad \Delta q_1(5) = \Delta q_2(1) = \Delta q_2(5) = 0 \]

\[ E(RC(6)) = E\{\Delta q_1(1)\Delta q_2(1) + \Delta q_1(2)\Delta q_2(2) + \Delta q_1(3)\Delta q_2(3) + \Delta q_1(4)\Delta q_2(4) + \Delta q_1(5)\Delta q_2(5) + \Delta q_1(6)\Delta q_2(6)\} \]

\[ = \int_{I_2} \omega_{12}(t)dt + \int_{I_3} \omega_{12}(t)dt + \int_{I_4} \omega_{12}(t)dt + \int_{I_6} \omega_{12}(t)dt \]
$M = 3$

\[ I_1 \quad I_2 \quad I_3 \]

\[ 0 \quad \frac{T}{3} \quad 2\frac{T}{3} \]

\[ \downarrow \]

$M = 6$

\[ I_2 \quad I_3 \quad I_4 \quad I_5 \quad I_6 \]

\[ 0 \quad \frac{T}{6} \quad \frac{2T}{6} \quad \frac{3T}{6} \quad \frac{4T}{6} \quad \frac{5T}{6} \quad T \]

‘The higher frequency, the smaller covered interval’
Lead & Lag modification

Bias-corrected estimator = Realized Covariance + Lead & Lag Terms

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

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

Simple lead-lag modification is not enough...⇒
the area covered by RC + Lead-Lag terms

We need another step before lead-lag modification⇒
First step (Modification for zero returns)

Price change vectors

\[(\Delta q_1(1), \Delta q_1(2), \Delta q_1(3), \Delta q_1(4), \Delta^2 q_1(6)) \]
\[(\Delta^2 q_2(2), \Delta q_2(3), \Delta q_2(4), \Delta^2 q_2(6)) \]

Contemporary part (sum up time-overlapping returns):

\[\Delta q_1(1)\Delta^2 q_2(2) + \Delta q_1(2)\Delta^2 q_2(2) + \Delta q_1(3)\Delta q_2(3) + \Delta q_1(4)\Delta q_2(4) + \Delta^2 q_1(6)\Delta^2 q_2(6).\]
Second step (lead-lag modification)

\[
(\Delta q_1(1), \Delta q_1(2), \Delta q_1(3), \Delta q_1(4), \Delta^2 q_1(6)) \\
(\Delta^2 q_2(2), \Delta q_2(3), \Delta q_2(4), \Delta^2 q_2(6))
\]

Lead-Lag part:

\[
\Delta q_1(2)\Delta q_2(3) \quad \Delta q_1(3)\Delta q_2(4) \quad \Delta q_1(4)\Delta^2 q_2(6) \\
\Delta^2 q_2(2)\Delta q_1(3) \quad \Delta q_2(3)\Delta q_1(4) \quad \Delta q_2(4)\Delta^2 q_1(6)
\]

Bias-corrected estimator:

\[BC(6) = \text{Contemporary part} + \text{Lead-Lag part}\]
Bias-corrected estimator

Similarly, in general case, the modification consists of two steps:
(1) zero-return modification (consider price change vectors and sum up products of time-overlapping changes)
(2) lead-lag modification (add products of lead-lag changes)

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

where

\[ m^-_i = \max\{m'_i < m_i : q_i(m'_i) \neq q_i(m'_i - 1)\}, \]
\[ A_i = \{q_i(m_i) \neq q_i(m_i - 1)\} \]
\[ A = \left\{ [m^-_1, m_1] \cap [m^-_2, m_2] \neq \emptyset \right\} \]
RC(M) and BC(M) for a large M

Hayashi and Yoshida (2005, Bernoulli)’s unbiased estimator for “transaction data”:

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

where \( B = \{(t_{n_1-1}, t_{n_1}] \cap (t_{n_2-1}, t_{n_2}] \neq \emptyset\} \).

For large \( M \)

\[ RC(M) = 0, \quad (1) \]
\[ BC(M) = HY. \quad (2) \]
Monte Carlo study

■ DGP:

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

\[d\sigma_{ij}(t) = \kappa \left( \theta - \sigma_{ij}(t) \right) dt + \gamma dz_{ij}(t), \quad i, j = 1, 2.\]

where \(\kappa = 0.01, \theta = 0.01,\) and \(\gamma = 0.001\) and \(T = 60 \times 60 \times 4.5\) seconds. \((p_1\) and \(p_2\) are positively correlated in average\)

■ Time differences are drawn from an exponential distribution:

\[F(t_{n_i} - t_{n_i-1}) = 1 - \exp \left\{ -\lambda_i (t_{n_i} - t_{n_i-1}) \right\}, \quad i = 1, 2\]

where \(F(\cdot)\) denotes a cumulative distribution function, \(\lambda_1 = 23.4\) and \(\lambda_2 = 20.9.\)

■ The number of replications is 500.
Monte Carlo result

Distribution of measurement error (500 replications)
An empirical example

Daily $RC$ and $BC$ (20 days, Honda-Nissan)
Summary

- An explanation of the cause of the bias on RC
- A bias-corrected estimator

Remaining works:
- Microstructure noise (Observation error)
- Evaluation of Variance of the estimator
  ⇒ MSE analysis