Consistent Realized Covariance for Asynchronous Observations Contaminated by Market Microstructure Noise

Alessandro Palandri*

September 2006

Abstract

This paper proposes a consistent estimator for the realized covariance of high frequency and asynchronous assets’ returns that are contaminated by microstructure noise. The main contribution is the introduction of the pseudo-aggregation which transforms the observations into series with the same number of data points without incurring in any loss of information. This in turn makes it possible to construct an unbiased and consistent covariance estimator by merging techniques from the literature relating to the asynchrony and the market microstructure contamination. Monte Carlo simulations confirm the theoretical results and highlight the outstanding performance of the proposed estimator.

Keywords: Realized Covariance, Asynchronous Observations, Market Microstructure, Sub-sampling, Pseudo-Aggregation.

JEL classification: C32, G0, G1.

*Department of Economics, University of Copenhagen, 408: Bispetorvet, 4. sal, DK-1455 Copenhagen K, Denmark, Email: alessandro.palandri@econ.ku.dk
1 Introduction

The analysis and modeling of second order moments of assets’ returns is fundamental in many financial problems such as portfolio management and selection, risk analysis and hedging and the pricing of assets and derivatives. The availability of high frequency data has led to the nonparametric estimation of the daily variance of such returns by *Realized Volatility*. In an idealized diffusion world, this estimator reduces to the sum of the squares of the high-frequency returns within a given day. Analogously, the daily covariance between two assets could be estimated by *Realized Covariance*, defined as the sum of the cross-products of high-frequency returns. Unfortunately, the non-simultaneity in the realizations of the returns are not simultaneous across assets causes the loss of the common notion of cross-products, where to every observation of a given asset corresponds one and only one observation of any other asset. The usual approach for dealing with this issue is to artificially synchronize the data by aggregating the returns over regularly spaced time intervals, thus constructing time-aggregated returns, for which it is then straightforward to define the necessary cross-products.

However, such method presents two drawbacks in terms of variance and bias of the estimator. The fact that, for a given number of realizations, a potentially large fraction of the data is being lost in the aggregation implies a non arguable loss of efficiency. Furthermore, for a fixed interval of time aggregation the variance of the estimator is insensitive to the sample size, making it impossible to achieve consistent estimates. The second drawback derives from the fact that the returns, which are observed at random times, will not fill the chosen time intervals properly. As it has been shown through simulations and empirical studies, this results in a bias of the estimator toward zero which increases with the reduction of the time span over which the returns are being aggregated. Hayashi and Yoshida (2005) and Corsi (2006) have independently proposed tick-by-tick covariance estimators that make use of all the available data and do not require any artificial synchronization of the observations. In turn their estimator, henceforth HYC, is consistent and unbiased under the assumption of continuous time stochastic processes observable at random times.

Although it is justified under the stated assumptions, in the presence of market microstructure (MM) noise the HYC estimator, while retaining unbiasedness, becomes inconsistent as its variance diverges linearly with the number of observations. A similar issue has been well documented and object of extensive study in the Re-
alized Volatility literature. There the $MM$ contamination causes the corresponding estimator to diverge since, as the sampling frequency goes to infinity, an infinite number of random variables with non zero mean and fixed variance enter the summation defining the realized variance estimator. In the HYC the drift to infinity is missing because the $MM$ noise is generally uncorrelated across assets returns, whether they are simultaneous or not. Nevertheless, asymptotically, an infinite number of zero mean random variables with fixed variance entering the summation is enough to cause the inconsistency of the estimator.

If the realizations of assets’ returns were simultaneous, defining the Realized Covariance estimator would be straightforward and the methods for dealing with the presence of $MM$ noise could be easily “imported” from the literature on Realized Volatility. In particular, it would be possible to employ the consistent estimator of Zhang, Mykland and Aït-Sahalia (2005), henceforth ZMA. Here, the $MM$ noise is effectively damped by constructing $K$ series of aggregate returns of $K$ ticks$^1$ which are then used to compute $K$ intermediate and inconsistent estimators$^2$ that will be averaged to obtain, at last, the desired consistent estimator.

It seems quite natural to expect a consistent Realized Covariance estimator for asynchronous observations contaminated by $MM$ noise to have many points in common with the HYC and the ZMA estimators. Specifically, it should mimic the treatment of the asynchrony of the HYC and the damping of the $MM$ contamination of the ZMA. The main contribution of this paper is the solution to the problem of linking the HYC and ZMA estimators and their statistical properties. Such link takes the form of the pseudo-aggregation of the tick data which, by eliminating useless observations$^3$, generates series of pseudo-aggregate returns with exactly the same number of points. It is then straightforward to extend the approach of Zhang, Mykland and Aït-Sahalia to the pseudo-aggregated data to derive the Consistent Realized Covariance (CRC) estimator.

The paper is organized as follows. Section 2 defines the concept of covariance for

---

$^1$The first series is constructed by aggregating every $K$ observations beginning from the first and up to the $N - K$, where $N$ is the total number of observations. The $n$-th series is constructed in the same way but starting from the $n$-th observation up to the $N - K + n - 1$.

$^2$Since each $MM$ error enters one and only one of the intermediate estimators, the average of such estimators shrinks the impacting variance from the $MM$ noise by a factor $K$.

$^3$In the framework of covariance estimation of asynchronous series, the paradigm stating to “never throw data away” only holds in the weaker form “never throw useful data away”. The distinction highlights the fact that when computing covariances the sample will generally contain observations that are useless and that can therefore be discarded without any concern. See example 3.2 supporting this statement.
asynchronous observations. Section 3 introduces the simple algorithm that generates
the pseudo-aggregate returns, necessary for the definition of a consistent covariance
estimator, together with a few examples showing its inner workings. Follows Sec-
tion 4 in which the Consistent Realized Covariance estimator is defined and where
the asymptotic properties are derived. The simulation results comparing the exist-
ing methodologies and the CRC estimator are presented in Section 5. Section 6
concludes.

2 Covariance Definition

For a pair of returns \( x \) and \( y \) that are observable only at the daily frequency, their
cross-product provides an unbiased but inconsistent estimator of the daily covariation:

\[
E[x \cdot y] = \sigma_{xy} \equiv \rho_{xy} \sigma_x \sigma_y
\]  

(1)

In the case of high-frequency asynchronous observations, the realized covariance es-
timator can be defined as:

\[
\hat{\sigma}_{xy} = \frac{1}{N_x \sum_{i=1}^{N_x} x_i y_j} \cdot \mathbb{1}[E(x_i y_j | \mathcal{M}) \neq 0]
\]  

(2)

where \( \mathcal{M} \) denotes the true data generating process and \( \mathbb{1}[\bullet] \) is the indicator function
which takes the value of one only when the corresponding cross-product of high-
frequency returns has non-zero expectation conditional on \( \mathcal{M} \). Notice how, replacing
the indicator function with the number 1, yields the covariance estimator of equation
(1). In equation (2), removing the cross-products with zero expectation under \( \mathcal{M} \) has
no impact on the expected value of the estimator but it does reduce its variance by
cutting back on the number of random variables entering the double summation:

\[
E[\hat{\sigma}_{xy}] = \sigma_{xy}
\]

\[
\mathbb{V}[\hat{\sigma}_{xy}] < \mathbb{V}[x \cdot y]
\]

Under mild assumptions about the data generating process \( \mathcal{M} \), the variance reduction
effect, induced by the indicator function, is so strong to result in the consistency of
the realized covariance estimator \( \hat{\sigma}_{xy} \).

In an ideal world without market microstructure contamination and under the
assumption that the high-frequency asynchronous returns are realizations of randomly
observable continuous processes, \( \mathcal{M} \) takes the following specification:

\[
    x_{t_i} = \sigma_x \int_{t_{i-1}}^{t_i} d B_x \\
    y_{\tau_j} = \rho_{xy} \sigma_y \int_{\tau_{j-1}}^{\tau_j} d B_x + (1 - \rho_{xy}^2)^{1/2} \sigma_y \int_{\tau_{j-1}}^{\tau_j} d B_y
\]

where \( B_x \) and \( B_y \) are standardized independent Brownian motions and \( t_i \) and \( \tau_j \) are the instants at which the returns \( x_i \) and \( y_j \) are being observed. Conditional on this specification of the data generating process, the expected value of a given cross-product is non-zero if and only if the corresponding time intervals overlap:

\[
    \mathbb{E} (x_{t_i} y_{\tau_j} | \mathcal{M}) \neq 0 \iff \min(t_i, \tau_j) > \max(t_{i-1}, \tau_{j-1})
\]

where \( t_i \) and \( \tau_j \) are the instants in which the returns \( x_i \) and \( y_j \) are observed. A root-\( N \), as defined in Section 3, consistent covariance estimator can then be defined as:

\[
    \hat{\sigma}_{xy} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} x_i y_j \cdot \mathbb{1} [\min(t_i, \tau_j) > \max(t_{i-1}, \tau_{j-1})]
\]

This estimator, proposed by Hayashi and Yoshida (2005) and Corsi (2006), has an asymptotic behavior which parallels that of the tick-by-tick realized variance estimator. That is, in the presence of market microstructure noise it is inconsistent and its variance increases linearly with the number of observations \( N \).

### 3 Pseudo-Aggregation

In equation 3 the indicator function takes the value one only when the time intervals of two observations \( x \) and \( y \) overlap. Since for any pair of assets, any given return may overlap with multiple returns, there will be cross-products sharing a common factor. This implies the existence of terms which can be grouped together. Among the possible aggregations, the pseudo-aggregation consists in the construction of the sets of observations \( \{a_i\}_{i=1}^N \) and \( \{b_i\}_{i=1}^N \) according to the following algorithm:
\[ \text{if } t_q < \tau_r \quad \text{if } t_q = \tau_r \quad \text{if } t_q > \tau_r \]

| \( a_i = \{x_{t_q}, \ldots, x_{t_q'}\} \) | \( a_i = \{x_{t_q}\} \) | \( a_i = \{x_{t_q}\} \) |
| \( b_i = \{y_{\tau_r}\} \) | \( b_i = \{y_{\tau_r}\} \) | \( b_i = \{y_{\tau_r}, \ldots, y_{\tau_r'}\} \) |

where \( t_{q'-1} < \tau_r \leq t_{q'} \)

\[
\begin{cases} 
\text{if } t_{q'} = \tau_r & q = q' + 1 \\
\text{else} & q = q' \\
r = r + 1 
\end{cases}
\]

\[
\begin{cases} 
\text{if } t_{q'} = \tau_r & q = q + 1 \\
\text{else} & q = q' \\
r = r + 1 
\end{cases}
\]

\[
\begin{cases} 
\text{if } t_q = \tau_{r'} & r = r' + 1 \\
\text{else} & r = r' 
\end{cases}
\]

\( i = i + 1 \) and back to top

The sets \( a_i \) and \( b_i \) allow to construct the pseudo-aggregate returns \( \tilde{x}_i \) and \( \tilde{y}_i \) for all \( i = 1, \ldots, N \):

\[
\begin{align*}
\tilde{x}_i &= \sum_{x \in a_i} x \\
\tilde{y}_i &= \sum_{y \in b_i} y
\end{align*}
\]

Since the sets \( a_i \) and \( b_i \) and the pseudo-aggregate returns \( \tilde{x}_i \) and \( \tilde{y}_i \), reflect one of the possible factorizations of the cross-products in the Hayashi-Yoshida-Corsi estimator of equation 3, it follows that:

\[
\sum_{i=1}^{N} \tilde{x}_i \tilde{y}_i = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} x_i y_j \cdot \mathbb{I} [\min (t_i, \tau_j) > \max (t_{i-1}, \tau_{j-1})]
\]

The pseudo-aggregation does not cause any loss of information because it only eliminates those observations that do not contribute any additional information to the determination of the covariance between the two series. Furthermore, while the number of observations of tick-by-tick returns \( N_x \) and \( N_y \) are generally different, the pseudo-aggregated returns comprise the same number of observations \( N \). This quantity is a good indicator of the real number of observations available and therefore a relevant quantity in the derivation of the asymptotic behavior of the estimator. The following examples show the inner working of the pseudo-aggregation and through simple and stylized cases provide some intuition for why the pseudo-sample size \( N \) is a good indicator of the information that can actually be used.
Example 3.1. Suppose to observe \( x \) and \( y \) simultaneously during the trading day:

\[
t : \quad x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad N_x = 5
\]

\[
\tau : \quad y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5 \quad N_y = 5
\]

\[
\begin{align*}
\circ x_1 &= x_1 \\ \circ x_2 &= x_2 \\ \circ x_3 &= x_3 \\ \circ x_4 &= x_4 \\ \circ x_5 &= x_5 \\
\circ y_1 &= y_1 \\ \circ y_2 &= y_2 \\ \circ y_3 &= y_3 \\ \circ y_4 &= y_4 \\ \circ y_5 &= y_5
\end{align*}
\]

\( N = 5 \)

Because of the natural synchronization of the data, the pseudo-aggregate returns coincide with the observations and \( N_x = N_y = N \) as expected.

Example 3.2. If both \( x \) and \( y \) are observed \( S > 2 \) times during the trading day but \( x \) is observed \( S - 1 \) times up to the middle of the day and once at the closing while \( y \) is observed once in the middle of the day and \( S - 1 \) times up to the closing:

\[
t : \quad x_1 \quad x_2 \quad \ldots \quad x_{S-1} \quad x_S \quad N_x = S
\]

\[
\tau : \quad y_1 \quad y_2 \quad \ldots \quad y_{S} \quad N_y = S
\]

\[
\begin{align*}
\circ x_1 &= x_1 + \cdots + x_{S-1} \\
\circ y_1 &= y_1 \\
\circ x_2 &= x_{S} \\
\circ y_2 &= y_2 + \cdots + y_{S}
\end{align*}
\]

\( N = 2 \)

The fact that 2 pseudo-aggregate returns have the same informational content of \( S > 2 \) returns is a direct consequence of the particular set of time stamps of this example. Fortunately, the structure of real data is hardly so extreme to reduce quite a large number of observations to only a few that are useful. Nevertheless, this type of effect arises whenever two series of returns are observed asynchronously.

Example 3.3. In the case in which \( x \) and \( y \) are observed asynchronously at random times:

\[
t : \quad x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7 \quad x_8 \quad N_x = 8
\]

\[
\tau : \quad y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5 \quad y_6 \quad N_y = 6
\]

\[
\begin{align*}
\circ x_1 &= x_1 \\
\circ y_1 &= y_1 + y_2 + y_3 \\
\circ x_2 &= x_2 \\
\circ y_2 &= y_3 + y_4 + y_5 \\
\circ x_3 &= x_3 + x_4 + x_5 \\
\circ y_3 &= y_5 \\
\circ x_4 &= x_5 + x_6 \\
\circ y_4 &= y_6
\end{align*}
\]

\( N = 4 \)

This case exemplifies the fact that \( N \) is a function of the specific times at which the returns are observed and together with the previous examples it should leave no doubt
about the fact that $N$ cannot be deduced from $N_x$ and $N_y$ alone, but that it is a function of the series of time stamps.

It is now possible to formally state the consistency of HYC covariance estimator in the absence of $MM$ noise in terms of the pseudo-aggregate returns:

$$\lim_{N \to \infty} \sum_{i=1}^{N} \circ x_i \circ y_i = \sigma_{xy}$$

$$\mathbb{V} \left[ \sum_{i=1}^{N} \circ x_i \circ y_i \right] \propto N^{-1}$$

However, when $MM$ noise is brought into the picture, the HYC estimator diverges with $N$:

$$\mathbb{V} \left[ \sum_{i=1}^{N} \circ x_i \circ y_i \right] \propto N$$

$$\lim_{N \to \infty} \mathbb{V} \left[ \sum_{i=1}^{N} \circ x_i \circ y_i \right] = +\infty$$

It should be noticed that $N$ does not coincide with the number of uncorrelated cross-products of pseudo-aggregate returns in the summation. However, the autocorrelations of the pseudo-aggregate returns are zero, by construction, after the second lag. This makes the number of uncorrelated cross-products no less than $N/3$ and therefore proportional to $N$.

4 Consistent Estimator

The sets of pseudo-aggregate returns $a_i$ and $b_i$ allow to derive a consistent estimator of the realized covariance in the presence of $MM$ noise by extending the approach proposed by Zhang, Mykland and Aït-Sahalia (2005) in the realized volatility framework. In particular, for given $K < N$, let’s define the sets $A_{i,j}$ and $B_{i,j}$, for $i = 1, ..., K$ and $j = 1, ..., N/K - 1$, given by the union of $K$ sets $a_i$ and $b_i$ respectively:

$$A_{i,j} = \bigcup_{l = (j-1)K + i}^{jK+(i-1)} a_l$$

and

$$B_{i,j} = \bigcup_{l = (j-1)K + i}^{jK+(i-1)} b_l$$

These sets allow to construct the aggregate returns $X_{i,j}$ and $Y_{i,j}$ of $K$ pseudo-aggregate returns:

$$X_{i,j} = \sum_{x \in A_{i,j}} x$$

and

$$Y_{i,j} = \sum_{y \in B_{i,j}} y$$
which parallel the aggregations in the ZMA variance estimator. The sets \( A_{i,j} \) and \( B_{i,j} \) cover all the possible ways of aggregating the sets of pseudo-aggregate returns \( a_l \) and \( b_l \) in groups of \( K \). These aggregate returns can then be used to compute \( K \) different covariance estimates:

\[
C_i = \omega_i \sum_{j=1}^{N/K-1} X_{i,j} Y_{i,j} \quad \text{with } i = 1, \ldots, K
\]

where the parameter \( \omega_i \) corrects for the fact that the summation only measures a fraction of the daily covariance as one cross-product of aggregate returns or \( K \) pseudo-aggregate returns are left out. These intermediate estimators are then used to construct the Consistent Realized Covariance estimator \( CRC \):

\[
CRC = \frac{1}{K} \sum_{i=1}^{K} C_i
\]

\[
\mathbb{E}[CRC] = \sigma_{xy} \equiv \rho_{xy} \sigma_x \sigma_y
\]

\[
\mathbb{V}[CRC] = \frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[C_p, C_q]
\]

Assuming that the observed log-price processes \( p_{t_i} \) are equal to the latent true log-price processes \( \hat{p}_{t_i} \) plus some i.i.d. noise, the observed aggregate return processes for assets \( x \) and \( y \) will be equal to:

\[
X_{i,j} = \hat{X}_{i,j} + \left( U_{t_{i,j}} - U_{\bar{t}_{i,j}} \right)
\]

\[
Y_{i,j} = \hat{Y}_{i,j} + \left( V_{\tau_{i,j}} - V_{\bar{\tau}_{i,j}} \right)
\]

where \( \hat{X}_{i,j} \) and \( \hat{Y}_{i,j} \) are the true latent aggregate returns, \( U \) and \( V \) are the microstructure errors with time indexes:

\[
\bar{t}_{i,j} = \sup \{ t : x_t \in A_{i,j} \} \quad , \quad \bar{\tau}_{i,j} = \sup \left[ t < \inf \{ t : x_t \in A_{i,j} \} : x_t \in \bigcup_{l=1}^{N} a_l \right]
\]

\[
\bar{t}_{i,j} = \sup \{ t : y_t \in B_{i,j} \} \quad , \quad \bar{\tau}_{i,j} = \sup \left[ \tau < \inf \{ \tau : y_t \in B_{i,j} \} : y_t \in \bigcup_{l=1}^{N} b_l \right]
\]

It follows that the cross products of the aggregate observable returns \( X_{i,j} Y_{i,j} \) can be written as:

\[
\hat{X}_{i,j} \hat{Y}_{i,j} + \hat{X}_{i,j} \left( V_{\tau_{i,j}} - V_{\bar{\tau}_{i,j}} \right) + \hat{Y}_{i,j} \left( U_{t_{i,j}} - U_{\bar{t}_{i,j}} \right) + \left( U_{t_{i,j}} - U_{\bar{t}_{i,j}} \right) \left( V_{\tau_{i,j}} - V_{\bar{\tau}_{i,j}} \right)
\]
allowing to rewrite the $C_i$ estimators as:

$$C_i = 1_i + 2_i + 3_i + 4_i$$

where:

$$1_i = \omega_i \sum_{j=1}^{N/K-1} \hat{X}_{i,j} \hat{Y}_{i,j}$$

$$2_i = \omega_i \sum_{j=1}^{N/K-1} \hat{X}_{i,j} \left( V_{\tau_{i,j}} - \hat{V}_{\tau_{i,j}} \right)$$

$$3_i = \omega_i \sum_{j=1}^{N/K-1} \hat{Y}_{i,j} \left( U_{t_{i,j}} - \hat{U}_{t_{i,j}} \right)$$

$$4_i = \omega_i \sum_{j=1}^{N/K-1} \left( U_{t_{i,j}} - \hat{U}_{t_{i,j}} \right) \left( V_{\tau_{i,j}} - \hat{V}_{\tau_{i,j}} \right)$$

By assumption the microstructure errors are uncorrelated with the true returns and across assets, thus:

$$\text{COV}[C_p, C_q] = \text{COV}[1_p, 1_q] + \text{COV}[2_p, 2_q] + \text{COV}[3_p, 3_q] + \text{COV}[4_p, 4_q]$$

making it possible to rewrite Equation (4) in the following expanded form:

$$\text{V}[CRC] = \frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[1_p, 1_q] + \frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[2_p, 2_q] + \frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[3_p, 3_q] + \frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[4_p, 4_q] \quad (5)$$

By construction, the pseudo-aggregate sets are such that $a_l \neq a_{l+m}$ and $b_l \neq b_{l+m}$ for every $|m| > 1$. This implies that the microstructure errors $V$ and $U$ entering respectively $2_p$ and $3_p$ are different from those entering $2_{p+m}$ and $3_{p+m}$ for every $|m| > 1$. Together with the assumption that the MM noise is serially uncorrelated, it follows that $\text{COV}[2_p, 2_{p+m}] = 0$ and $\text{COV}[3_p, 3_{p+m}] = 0$ for every $|m| > 1$. 

10
Thus:

\[
\frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[\mathcal{2}_p, \mathcal{2}_q] = \frac{1}{K^2} \sum_{p=1}^{K} \mathcal{V}[\mathcal{2}_p] + \frac{2}{K^2} \sum_{p=1}^{K-1} \text{COV}[\mathcal{2}_p, \mathcal{2}_{p+1}]
\]

\[
\frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[\mathcal{3}_p, \mathcal{3}_q] = \frac{1}{K^2} \sum_{p=1}^{K} \mathcal{V}[\mathcal{3}_p] + \frac{2}{K^2} \sum_{p=1}^{K-1} \text{COV}[\mathcal{3}_p, \mathcal{3}_{p+1}]
\]

The fact that, by construction, there can be pseudo-aggregate sets \(a_l = a_{l+1} \neq a_{l+2}\) and \(b_l = b_{l+1} \neq b_{l+2}\) is the sole reason why the pairs \(\mathcal{2}_p, \mathcal{2}_{p+1}\) and \(\mathcal{3}_p, \mathcal{3}_{p+1}\) will in general be correlated. Since the covariance is bounded above by the variance, there exist the constants \(\lambda_2\) and \(\lambda_3\), that do not depend neither on \(N\) nor \(K\), such that:

\[
\frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[\mathcal{2}_p, \mathcal{2}_q] = \frac{\lambda_2}{K^2} \sum_{p=1}^{K} \mathcal{V}[\mathcal{2}_p]
\]

\[
\frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[\mathcal{3}_p, \mathcal{3}_q] = \frac{\lambda_3}{K^2} \sum_{p=1}^{K} \mathcal{V}[\mathcal{3}_p]
\]

Furthermore, since the unconditional variances of the terms \(\mathcal{2}_p\) and \(\mathcal{3}_p\) do not depend on the index \(p\):

\[
\frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[\mathcal{2}_p, \mathcal{2}_q] = \frac{\lambda_2}{K^2} \cdot K \cdot \mathcal{V}[\mathcal{2}_p] = \lambda_2 \cdot \mathcal{V}[\mathcal{2}_p]
\]

\[
\frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[\mathcal{3}_p, \mathcal{3}_q] = \frac{\lambda_3}{K} \cdot \mathcal{V}[\mathcal{3}_p]
\]

These results apply *mutatis mutandis* to the covariances of \(\mathcal{4}_p\) and \(\mathcal{4}_q\):

\[
\frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[\mathcal{4}_p, \mathcal{4}_q] = \frac{\lambda_4}{K} \cdot \mathcal{V}[\mathcal{4}_p]
\]

With respect to the covariances of the terms \(\mathcal{1}_p\) and \(\mathcal{1}_q\), however, things are slightly different. On one hand, the uncorrelated \(MM\) errors which guarantee that the covariances are zero between the terms with \(|p - q| > 1\) are not present, on the other
the sums of the cross-products of aggregate returns have a non zero covariance for every \( p \) and \( q \). Hence:

\[
\frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \text{COV}[\mathbb{1}_p, \mathbb{1}_q] = \frac{1}{K} \sum_{p=1}^{K} \frac{1}{K} \sum_{q=1}^{K} \lambda_{1,p,q} \mathbb{V}[\mathbb{1}_p] \\
= \frac{1}{K} \sum_{p=1}^{K} \lambda_{1,p} \mathbb{V}[\mathbb{1}_p] \\
= \lambda_1 \cdot \mathbb{V}[\mathbb{1}_p]
\]

where \( \lambda_{1,p,q} \) is a constant that rescales the variance of \( \mathbb{1}_p \) to match the covariance between \( \mathbb{1}_p \) and \( \mathbb{1}_q \), \( \lambda_{1,p} \) is the sample average w.r.t. the index \( q \) of the \( K \) constants \( \lambda_{1,p,q} \) and \( \lambda_1 \) is the sample average of \( \lambda_{1,p} \). The partial results obtained for the double summations of covariances permit to rewrite Equation (5) as:

\[
\mathbb{V}[CRC] = \lambda_1 \cdot \mathbb{V}[\mathbb{1}_p] + \frac{\lambda_2}{K} \cdot \mathbb{V}[\mathbb{2}_p] + \frac{\lambda_3}{K} \cdot \mathbb{V}[\mathbb{3}_p] + \frac{\lambda_4}{K} \cdot \mathbb{V}[\mathbb{4}_p] \tag{6}
\]

By definition, the variance of the element \( \mathbb{4}_p \) is equal to:

\[
\mathbb{V}[\mathbb{4}_p] = \mathbb{V} \left[ N/K - \frac{1}{N} \sum_{j=1}^{N/K-1} (U_{t,j} - U_{\mu,j}) \left( V_{\tau,j} - V_{\mu,j} \right) \right] \tag{7}
\]

In order to make the notation more accessible, let’s define the following quantities:

\[
\Delta U_{t,p,j} = (U_{t,j} - U_{\mu,j}) \\
\Delta V_{\tau,p,j} = (V_{\tau,j} - V_{\mu,j})
\]

and substitute them in Equation (7):

\[
\mathbb{V}[\mathbb{4}_p] = \mathbb{V} \left[ \omega_p \sum_{j=1}^{N/K-1} \Delta U_{t,p,j} \cdot \Delta V_{\tau,p,j} \right] \\
= \mathbb{E} \left[ \omega_p^2 \right] \cdot \mathbb{E} \left[ \left( \sum_{j=1}^{N/K-1} \Delta U_{t,p,j} \cdot \Delta V_{\tau,p,j} \right)^2 \right] \\
- \left( \mathbb{E} [\omega_p] \right)^2 \cdot \mathbb{E} \left[ \sum_{j=1}^{N/K-1} \Delta U_{t,p,j} \cdot \Delta V_{\tau,p,j} \right]^2
\]
and since the MM errors $U$ and $V$ are uncorrelated and have zero mean:

$$
\mathbb{V}[(\mathbb{1})_p] = \mathbb{E}[\omega^2_p] \cdot \mathbb{V} \left[ \sum_{j=1}^{N/K-1} \Delta U_{tp,j} \cdot \Delta V_{tp,j} \right]
$$

$$
= \mathbb{E}[\omega^2_p] \cdot \sum_{j=1}^{N/K-1} \sum_{l=1}^{N/K-2} \text{COV} \left[ \Delta U_{tp,j} \cdot \Delta V_{tp,j}, \Delta U_{tp,j+1} \cdot \Delta V_{tp,j+1} \right]
$$

As previously stated, aggregate returns that are more than one lag apart do not share microstructure errors:

$$
\mathbb{V}[(\mathbb{1})_p] = \mathbb{E}[\omega^2_p] \cdot \sum_{j=1}^{N/K-1} \mathbb{V} \left[ \Delta U_{tp,j} \cdot \Delta V_{tp,j} \right]
$$

$$
+ 2\mathbb{E}[\omega^2_p] \cdot \sum_{j=1}^{N/K-1} \text{COV} \left[ \Delta U_{tp,j} \cdot \Delta V_{tp,j}, \Delta U_{tp,j+1} \cdot \Delta V_{tp,j+1} \right]
$$

$$
= \mathbb{E}[\omega^2_p] \cdot \phi_4 \sum_{j=1}^{N/K-1} \mathbb{V} \left[ \Delta U_{tp,j} \cdot \Delta V_{tp,j} \right]
$$

where $\phi_4$ is a constant that rescales the sum of the variances to incorporate the sum of the lag-one covariances. Using once more the fact that the microstructure errors $U$ and $V$ are uncorrelated and have mean zero:

$$
\mathbb{V}[(\mathbb{1})_p] = \mathbb{E}[\omega^2_p] \cdot \phi_4 \sum_{j=1}^{N/K-1} \mathbb{E} \left[ \Delta U_{tp,j}^2 \right] \cdot \mathbb{E} \left[ \Delta V_{tp,j}^2 \right]
$$

$$
= \mathbb{E}[\omega^2_p] \cdot \phi_4 \sum_{j=1}^{N/K-1} 2S_U^2 \cdot 2S_V^2
$$

$$
= 4 \cdot \mathbb{E}[\omega^2_p] \cdot \phi_4 \cdot S_U^2 S_V^2 \left( \frac{N}{K} - 1 \right)
$$

where $S_U^2$ and $S_V^2$ are the variances, respectively, of the microstructure errors $U$ and $V$. Therefore, the quantity that enters the expression for the variance of the Consistent Realized Covariance estimator, can be rewritten as:

$$
\frac{\lambda_4}{K} \cdot \mathbb{V}[(\mathbb{1})_p] = \frac{\lambda_4}{K} \cdot \mathbb{E}[\omega^2_p] \cdot \phi_4 S_U^2 S_V^2 \left( \frac{N}{K} - 1 \right)
$$

$$
= \psi_4 \left( \frac{N}{K^2} - \frac{1}{K} \right)
$$

where $\psi_4$ incorporates all the constants of proportionality that do not depend neither on $N$ nor $K$. 

13
The variance of \( \mathcal{O}_p \), given that returns that are more than one lag apart do not share microstructure errors, is equal to:

\[
\mathbb{V} [ \mathcal{O}_p ] = \mathbb{V} \left[ \sum_{j=1}^{N/K-1} \mathbb{V} \left[ \hat{X}_{p,j} \Delta V_{p,j} \right] \right]
\]

\[
= \mathbb{E} \left[ \omega_p^2 \right] \sum_{j=1}^{N/K-1} \sum_{l=1}^{N/K-1} \text{COV} \left[ \hat{X}_{p,j} \Delta V_{p,j}, \hat{X}_{p,l} \Delta V_{p,l} \right]
\]

\[
= \mathbb{E} \left[ \omega_p^2 \right] \sum_{j=1}^{N/K-1} \mathbb{V} \left[ \hat{X}_{p,j} \Delta V_{p,j} \right]
\]

\[
+ 2 \mathbb{E} \left[ \omega_p^2 \right] \sum_{j=1}^{N/K-2} \text{COV} \left[ \hat{X}_{p,j} \Delta V_{p,j}, \hat{X}_{p,j+1} \Delta V_{p,j+1} \right]
\]

\[
= \mathbb{E} \left[ \omega_p^2 \right] \cdot \phi^2 \sum_{j=1}^{N/K-1} \mathbb{V} \left[ \hat{X}_{p,j} \Delta V_{p,j} \right]
\]

where \( \phi^2 \) is a constant that rescales the sum of the variances to incorporate the sum of the lag-one covariances. Since the true latent returns and the microstructure errors are uncorrelated and have zero mean:

\[
\mathbb{V} [ \mathcal{O}_p ] = \mathbb{E} \left[ \omega_p^2 \right] \cdot \phi^2 \sum_{j=1}^{N/K-1} \mathbb{E} \left[ \hat{X}_{p,j}^2 \right] \cdot \mathbb{E} \left[ \Delta V_{p,j}^2 \right]
\]

\[
= 2 \mathbb{E} \left[ \omega_p^2 \right] \phi^2 S^2_Y \sum_{j=1}^{N/K-1} \mathbb{E} \left[ \hat{X}_{p,j}^2 \right]
\]

\[
= 2 \mathbb{E} \left[ \omega_p^2 \right] \phi^2 S^2_Y \cdot \phi^2 \sigma^2_X
\]

where \( \phi^2 \) is a constant that rescales the variance over the whole period \( \sigma^2_X \) to account for the fact that in the summation one of the \( N/K \) aggregate returns has been excluded and that some of the raw returns are present in more than one aggregate return. Therefore, the quantity that enters the expression of the variance of the Consistent Realized Covariance estimator, can be rewritten as:

\[
\frac{\lambda_3}{K} \cdot \mathbb{V} [ \mathcal{O}_p ] = \frac{\lambda_2}{K} \cdot 2 \mathbb{E} \left[ \omega_p^2 \right] \phi^2 S^2_Y \cdot \phi^2 \sigma^2_X
\]

\[
= \psi_3 \cdot \frac{1}{K}
\]

where \( \psi_3 \) incorporates all the constants of proportionality that do no depend on \( K \).

This result applies mutatis mutandis to the variance of the \( \mathcal{O}_p \) term:

\[
\frac{\lambda_3}{K} \cdot \mathbb{V} [ \mathcal{O}_p ] = \psi_3 \cdot \frac{1}{K}
\]
Hence, the variance of the Consistent Realized Covariance estimator can be further simplified from Equation (6) to:

$$V[CRC] = \lambda_1 \cdot V[\hat{\Omega}_p] + \psi_2 \frac{1}{K} + \psi_3 \frac{1}{K} + \psi_4 \left( \frac{N}{K^2} - \frac{1}{K} \right)$$

$$= \lambda_1 \cdot V[\hat{\Omega}_p] + \frac{1}{K} (\psi_2 + \psi_3 - \psi_4) + \frac{N}{K^2} \psi_4$$  \hspace{1cm} (8)

The variance of the $\hat{\Omega}_p$ element is equal to:

$$V[\hat{\Omega}_p] = \mathbb{E} \left[ \omega_p^2 \right] \cdot \sum_{j=1}^{N/K-1} \sum_{l=1}^{N/K-1} \text{COV} \left[ \hat{X}_{p,j} \hat{Y}_{p,j}, \hat{X}_{p,l} \hat{Y}_{p,l} \right]$$

Here the structure of the covariances is slightly different from that encountered for the elements $\hat{\Omega}_p$, $\hat{\Omega}_p$, and $\hat{\Omega}_p$. By construction of the pseudo-aggregate sets $a_t$ and $b_l$, there can be cases in which two and no more than two adjacent true returns are in fact the same. In particular, consider the following case:

$$\hat{X}_{p,j} \hat{Y}_{p,j}, \hat{X}_{p,j} \hat{Y}_{p,j+1}, \hat{X}_{p,j+1} \hat{Y}_{p,j+1}, \hat{X}_{p,j+1} \hat{Y}_{p,j+2}$$

highlighting the fact that only cross-products that are more than two lags apart are certain to have zero covariance. Thus:

$$V[\hat{\Omega}_p] = \mathbb{E} \left[ \omega_p^2 \right] \cdot \sum_{j=1}^{N/K-1} V \left[ \hat{X}_{p,j} \hat{Y}_{p,j} \right]$$

$$+ 2 \mathbb{E} \left[ \omega_p^2 \right] \cdot \sum_{j=1}^{N/K-2} \sum_{l=j+1}^{N/K-2} \text{COV} \left[ \hat{X}_{p,j} \hat{Y}_{p,j}, \hat{X}_{p,l} \hat{Y}_{p,l} \right]$$

$$+ 2 \mathbb{E} \left[ \omega_p^2 \right] \cdot \sum_{j=1}^{N/K-3} \sum_{l=j+2}^{N/K-3} \text{COV} \left[ \hat{X}_{p,j} \hat{Y}_{p,j}, \hat{X}_{p,l} \hat{Y}_{p,l} \right]$$

$$= \mathbb{E} \left[ \omega_p^2 \right] \cdot \phi_1 \sum_{j=1}^{N/K-1} V \left[ \hat{X}_{p,j} \hat{Y}_{p,j} \right]$$  \hspace{1cm} (9)

where $\phi_1$ is the constant that rescales the sum of variances to incorporate the sum of lag-one and lag-two covariances. From the definition of the pseudo-aggregate returns, the time intervals will either be $t_{p,j} \leq T_{p,j} \leq \tau_{p,j}$ or $\tau_{p,j} \leq t_{p,j} \leq \tau_{p,j} \leq \bar{t}_{p,j}$.
This means that each aggregate return can be expressed as the sum of two parts, one overlapping $O$ and one not overlapping $NO$:

\[
\begin{align*}
\hat{X}_{p,j} &= O_{x,p,j} + NO_{x,p,j} \\
\hat{Y}_{p,j} &= O_{y,p,j} + NO_{y,p,j}
\end{align*}
\]

The variance of the cross-product of aggregate returns can then be rewritten as:

\[
V \left[ \hat{X}_{p,j} \hat{Y}_{p,j} \right] = V \left[ O_xO_y + O_xNO_y + NO_xO_y + NO_xNO_y \right]
\]

where for simplicity of notation, the indexes $p$ and $j$ have been dropped from both $O$ and $NO$. Since the non-overlapping terms $NO$ are uncorrelated with all the other terms and they have zero mean:

\[
V \left[ \hat{X}_{p,j} \hat{Y}_{p,j} \right] = V \left[ O_xO_y \right] + V \left[ O_xNO_y \right] + V \left[ NO_xO_y \right] + V \left[ NO_xNO_y \right]
\]

Without loss of generality, let:

\[
O_x = \sigma_x \int_\theta dB_x \\
O_y = \rho_{xy} \sigma_y \int_\theta dB_x + (1 - \rho_{xy}^2)^{1/2} \sigma_y \int_\theta dB_y
\]

which implies that the variances of the overlaps and non-overlaps are equal to the daily variance of the corresponding asset rescaled by their time span. Letting $\mathcal{O}$ and $\mathcal{NO}$ indicate the time spans respectively of $O$ and $NO$:

\[
V \left[ \hat{X}_{p,j} \hat{Y}_{p,j} \right] = V \left[ O_xO_y \right] + \sigma_x^2 \sigma_y^2 \left( E \left[ \mathcal{O} \right] E \left[ \mathcal{NO}_y \right] + E \left[ \mathcal{NO}_x \right] E \left[ \mathcal{O} \right] + E \left[ \mathcal{NO}_x \right] E \left[ \mathcal{NO}_y \right] \right)
\]

The variance of the product of the overlapping terms is equal to:

\[
V \left[ O_xO_y \right] = E \left[ O_x^2 O_y^2 \right] - \left( E \left[ O_xO_y \right] \right)^2
\]

\[
= E \left[ E \left[ O_x^2 O_y^2 \right] | \mathcal{O} \right] - \left( E \left[ O_xO_y \right] | \mathcal{O} \right)^2
\]

\[
= \rho_{xy}^2 \sigma_x^2 \sigma_y^2 E \left[ \left( \int_\theta dB_x \right)^4 \left| \mathcal{O} \right| \right]
\]

\[
+ (1 - \rho_{xy}^2) \sigma_x^2 \sigma_y^2 E \left[ \left( \int_\theta dB_x \right)^2 \left( \int_\theta dB_y \right)^2 \left| \mathcal{O} \right| \right] - \rho_{xy}^2 \sigma_x^2 \sigma_y^2 E \left[ \mathcal{O} \right]^2
\]
By the Central Limit Theorem \( \mathcal{O}^{-1/2} \int_0 \, dB \overset{d}{\rightarrow} N(0,1) \), the expected value of the integrals simplifies to:

\[
\mathbb{V} \left[ O_x O_y \right] = 3 \rho_{xy}^2 \sigma_x^2 \sigma_y^2 \mathbb{E} \left[ \mathcal{O}^2 \right] + \left( 1 - \rho_{xy}^2 \right) \sigma_x^2 \sigma_y^2 \mathbb{E} \left[ \mathcal{O}^2 \right] - \rho_{xy}^2 \sigma_x^2 \sigma_y^2 \mathbb{E} \left[ \mathcal{O} \right]^2
\]

\[
= \left( 1 + 2 \rho_{xy}^2 \right) \sigma_x^2 \sigma_y^2 \mathbb{E} \left[ \mathcal{O}^2 \right] - \rho_{xy}^2 \sigma_x^2 \sigma_y^2 \mathbb{E} \left[ \mathcal{O} \right]^2
\]

and therefore for the variance of the aggregate cross-products of Equation (10) becomes:

\[
\mathbb{V} \left[ \hat{X}_{p,j} \hat{Y}_{p,j} \right] = \sigma_x^2 \sigma_y^2 \left\{ \left( 1 + 2 \rho_{xy}^2 \right) \mathbb{E} \left[ \mathcal{O}^2 \right] - \rho_{xy}^2 \mathbb{E} \left[ \mathcal{O} \right]^2 \right\}
\]

\[
+ \mathbb{E} \left[ \mathcal{O} \right] \mathbb{E} \left[ \mathcal{O} \mathcal{O}_y \right] + \mathbb{E} \left[ \mathcal{O} \mathcal{O}_x \right] \mathbb{E} \left[ \mathcal{O} \right] + \mathbb{E} \left[ \mathcal{O} \mathcal{O}_x \right] \mathbb{E} \left[ \mathcal{O} \mathcal{O}_y \right]
\]

\[
= \sigma_x^2 \sigma_y^2 \left\{ \left( 1 + 2 \rho_{xy}^2 \right) \mathbb{V} \left[ \mathcal{O} \right] + \left( 1 + \rho_{xy}^2 \right) \mathbb{E} \left[ \mathcal{O} \right]^2 \right\}
\]

\[
+ \mathbb{E} \left( \tilde{t} - \bar{t} \right) \mathbb{E} \left[ \mathcal{O} \mathcal{O}_y \right] + \mathbb{E} \left( \tilde{t} - \bar{t} \right) \mathbb{E} \left[ \mathcal{O} \mathcal{O}_x \right]
\]

\[
= \sigma_x^2 \sigma_y^2 \left\{ \left( 1 + 2 \rho_{xy}^2 \right) \mathbb{V} \left[ \mathcal{O} \right] + \rho_{xy}^2 \mathbb{E} \left[ \mathcal{O} \right]^2 \right\}
\]

\[
+ \mathbb{E} \left( \tilde{t} - \bar{t} \right) \mathbb{E} \left[ \mathcal{O} \mathcal{O}_x \right] + \mathbb{E} \left( \tilde{t} - \bar{t} \right) \mathbb{E} \left( \tau - \bar{\tau} \right)
\]

From the fill-in asymptotics, when the number of pseudo-aggregate returns \( N \) doubles, the size of the time intervals halves. In the same fashion if the number \( K \), over which the aggregate returns are constructed, doubles, their corresponding time intervals will double too. Hence:

\[
\left( \tilde{t} - \bar{t} \right) = \frac{K}{N} \cdot \alpha
\]

\[
\left( \tau - \bar{\tau} \right) = \frac{K}{N} \cdot \beta
\]

where \( \alpha \sim (\mu_\alpha, \sigma_\alpha^2) \) and \( \beta \sim (\mu_\beta, \sigma_\beta^2) \) are random variables. Since the number of overlaps is \( N/K - 1 \):

\[
\mathcal{O} = \frac{K}{N} \gamma
\]

with \( \gamma \) being a random variable with mean \( \mu_\gamma \) and variance \( \sigma_\gamma^2 \). The variance of the aggregate cross-products can then take its definitive expression:

\[
\mathbb{V} \left[ \hat{X}_{p,j} \hat{Y}_{p,j} \right] = \frac{K^2}{N^2} \sigma_x^2 \sigma_y^2 \left\{ \left( 1 + 2 \rho_{xy}^2 \right) \sigma_\gamma^2 + \rho_{xy}^2 \mu_\gamma^2 + \mu_\alpha \mu_\beta \right\}
\]

Substituting this expression in Equation (9) gives:

\[
\mathbb{V} \left[ \hat{y}_{p} \right] = \mathbb{E} \left[ \sigma_y^2 \right] \cdot \phi_1 \frac{K}{N} \sigma_x^2 \sigma_y^2 \left\{ \left( 1 + 2 \rho_{xy}^2 \right) \sigma_\gamma^2 + \rho_{xy}^2 \mu_\gamma^2 + \mu_\alpha \mu_\beta \right\}
\]

\[
= \psi_1 \frac{K}{N} 11 \quad (11)
\]
where \( \psi_1 \), as in the previous cases, incorporates all the constants of proportionality that do not depend neither on \( N \) nor \( K \). Finally, substituting Equation (11) in Equation (8) gives the variance of the Consistent Realized Covariance estimator:

\[
V[\text{CRC}] = \frac{K}{N} \psi_1 + \frac{1}{K} (\psi_2 + \psi_3 - \psi_4) + \frac{N}{K^2} \psi_4
\]

Since \( K \) is of an order less than \( N \), \( 1/K \) is of smaller order than \( N/K^2 \):

\[
V[\text{CRC}] = \frac{K}{N} \psi_1 + \frac{N}{K^2} \psi_4 + o \left( \frac{N}{K^2} \right)
\]

The size of \( K \) that minimizes the variance of the estimator \( CRC \) is the one for which the two elements of its variance are of the same order:

\[
\frac{K}{N} \propto \frac{N}{K^2} \quad \frac{K^3}{N} \propto \frac{N^2}{K^2}
\]

This, in turn, makes the size of the variance proportional to:

\[
V[\text{CRC}] \propto N^{-1/3}
\]

which not only completes the proof of the consistency of the CRC estimator, but it also shows how, in agreement with the results of Zhang, Mykland and Aït-Sahalia for the Realized Volatility estimator, convergence is at the rate \( N^{1/6} \).

5 Simulations

5.1 Simulations Set-Up

In this section the performances of the \( CRC \) and other competing estimators are evaluated through Monte Carlo simulations. The latent unobservable prices are generated by the following continuous time model:

\[
\hat{P}_{x,t_n} = \hat{P}_{x,t_{n-1}} + \sigma_x \int_{t_{n-1}}^{t_n} dB_x
\]

\[
\hat{P}_{y,\tau_n} = \hat{P}_{y,\tau_{n-1}} + \rho_{xy} \sigma_y \int_{\tau_{n-1}}^{\tau_n} dB_x + (1 - \rho_{xy}^2)^{1/2} \sigma_y \int_{\tau_{n-1}}^{\tau_n} dB_y
\]

where \( B_x \) and \( B_y \) are standardized independent Brownian motions and the observations time are independent random variables with exponential increments:

\[
t_n = t_{n-1} + \lambda \exp(1)
\]

\[
\tau_n = \tau_{n-1} + \lambda \exp(1)
\]
The simulated observable prices $P_{x,t_n}$ and $P_{y,\tau_n}$ are then generated by adding idiosyncratic noise representing the market microstructure:

$$P_{x,t_n} = \hat{P}_{x,t_n} + U_{t_n}$$
$$P_{y,\tau_n} = \hat{P}_{y,\tau_n} + V_{\tau_n}$$

with:

$$U_{t_n} \sim N(0, S_U^2) \quad \text{and} \quad V_{\tau_n} \sim N(0, S_V^2)$$

where the parameter $\eta$ is the noise to signal ratio. In the simulations, each trading day has been set to 7.5 hours corresponding to 27,000 seconds. The average waiting time between observations $\lambda$, measured in seconds, was allowed to take the following values $\{1, 2, 5, 15, 30, 60, 150, 300\}$. This range covers a wide range of assets going from the very liquid, traded once per second on average, to those with an average trading of up to five minutes. The noise to signal ratio was varied from 0.0004 to 0.0064 in multiples of two. Such values are consistent with the findings on real data of Hansen and Lunde (2005). The daily variance has been set equal to unity for both assets while the correlation has been varied from 0.0 to 0.5 in steps of 0.1. The total number of simulations, for each of the 240 combinations, was set to 10,000.

While the computation of the constant of proportionality for the optimal $K$ is too complicated to be carried out analytically, there are relevant factors that can be pulled out easily. Particularly, leaving out the $o(N/K^2)$ term, the first best choice of $K$ is such that:

$$K_{\text{opt}}^3 = 2 \cdot \frac{\psi_4}{\psi_1} \cdot N^2$$

$$= \text{constant} \cdot 8 \cdot \frac{S_U^2}{\sigma_x^2} \cdot \frac{S_V^2}{\sigma_y^2} \cdot N^2$$

A second best alternative to the optimal $K$ is to simply neglect the constant of proportionality and let the size of the aggregation window depend on the pseudo-aggregate sample size and the noise to signal ratios:

$$K_{\text{s.b.}} = 2 \left[ \left( \frac{S_U^2}{\sigma_x^2} \cdot \frac{S_V^2}{\sigma_y^2} \cdot N^2 \right)^{1/3} \right]$$

Finally, a third best alternative is to let the daily variances $\sigma_x^2$ and $\sigma_y^2$ be absorbed into the constant of proportionality thus eliminating the need to estimate them:

$$K_{\text{t.b.}} = 2 \left[ \left( S_U^2 \cdot S_V^2 \cdot N^2 \right)^{1/3} \right]$$
5.2 Simulations Results

The simulation results for the CRC estimator are compared to those of the HYC estimator and those obtained from 5, 10 and 15 minutes aggregated data. Figure 1 shows the bias of the five estimators for the median noise to signal ratio of 0.0016 and various values of the covariance and sampling frequency. Both the CRC and HYC estimators are unbiased in every simulated scenario. On the other hand, the estimators based on time-aggregated data show different levels of bias. For all of them such bias is toward zero and higher for the 5 minutes aggregate data and lower for the 15 minutes. As the lower half of the plots show, when the sampling frequency is high relative to the time aggregation, the bias is negligible for all three estimators. Yet, the first half of the plots show that when it is the time aggregation to be fine relative to the sampling frequency the magnitude of the bias can be very large. The log-MSE of the estimators are compared in Figure 2. The graphs do not show significant dependence of the log-MSE on the value of the covariance with the exception of the first two figures for which the upward sloping behavior is due to the substantial bias component. For low sampling frequencies, the different estimators are almost equivalent. However, as the available data becomes finer and finer, the log-MSE of the HYC estimator increases, that of the 5, 10 and 15 minutes aggregate data remains constant while the CRC is the only one to exhibit a decreasing log-MSE.

In Figure 3 the bias of the various estimators, for the mean value of the covariances, is plotted against the different values of the noise to signal ratio. The results are analogous to those of Figure 1: unbiasedness of the HYC and CRC estimators, considerable bias of the 5, 10 and 15 minutes estimators when the sampling frequency is relatively coarse and not surprisingly the independence of the bias from the level of microstructure noise. Figure 4 depicts the effects of increasing levels of $MM$ noise on the log-MSE of the estimators at different sampling frequencies. When the average spacing between the each tick is 300 seconds all five estimators are equivalent in terms of MSE. Nonetheless, an increase in the sampling frequency results in a clear separation of their performances. While higher noise to signal ratios have a mild effect on the estimators based on the time aggregation, they are deleterious for the HYC whose log-MSE increases linearly with the magnitude of the variance of the $MM$ noise. The log-MSE of the CRC estimator increases linearly as well but it does so at a very low rate which in terms of MSE translates into a smaller order of magnitude. Furthermore, at high frequencies, this estimator is clearly the most efficient as its log-MSE is only a fraction of that of the second best alternative corresponding to the
estimator based on 5 minutes aggregate data.

The biases as functions of the sampling frequency, for all the levels of the noise to signal ratio, are plotted in Figure 5. Again, while the HYC and CRC estimators are unbiased, the 5, 10 and 15 minutes estimators exhibit a bias toward zero that increases with the ratio of the sampling frequency to the time aggregation. In Figure 6 the plots of the log-MSE for the 5, 10 and 15 minutes estimators become flat quite rapidly. This is due to the fact that, by looking at a fixed time interval, these estimators literally throw away finer observations. Hence, their inconsistency. On the other hand, for low sampling frequencies the HYC and CRC estimators are equivalent in terms of log-MSE. Since in the absence of MM noise, the HYC estimator is the optimal choice, for relatively small noise to signal ratios and/or relatively low sampling frequencies, HYC is quasi-optimal. In such cases the aggregation parameter $K$ of the CRC estimator will be close to unity making it coincide with the HYC. Nevertheless, for significant noise to signal ratios and/or high sampling frequencies, the parameter $K$ takes values that are larger than unity and the different behavior of the CRC with respect to the HYC is evident. The log-MSE of the former decreases linearly with the sampling frequency while that of the latter increases linearly. Moreover, the MSE of the CRC estimator is not only the smallest but, as predicted by the theory, of a totally different order of magnitude. These figures highlight the inconsistency of the HYC estimator (its variance diverges asymptotically) and the 5, 10 and 15 minutes estimators (their variance stays constant) and the consistency of the CRC estimator (its variance converges to zero asymptotically).

The improvements of the CRC with respect to the other competing estimators, in terms of MSE, are reported in Figure 7. Here, for any given noise to signal ratio, when the average sampling frequency is 300 seconds the CRC estimator is outperformed by the 5, 10 and 15 minutes estimators. This follows from the fact that, under these circumstances, these estimators can in fact be seen as shrinkage estimators with a small variance but a large bias. In all the other cases the improvement in term of MSE for the CRC estimator with respect to the time aggregated estimators is strikingly, ranging from 20% up to more than 90%. The HYC estimator, on the other hand, never outperforms the CRC except in the few cases in which $K$ is equal to unity and for which the two estimators coincide. However, when the noise to signal ratio is non-negligible and/or the sampling frequency is not low, the MSE of the CRC estimator is from 40% up to 99% smaller than that of the HYC. While theoretically these results might not be surprising given that the CRC estimator is the only one for
which the variance converges to zero asymptotically, it is extremely important from an empirical point of view that for realistic values of the noise to signal ratio and sampling frequencies the gains deriving from the use of the proposed estimator are of such magnitude.

6 Conclusion

This paper develops a consistent covariance estimator for asynchronous high frequency data contaminated by market microstructure noise. The CRC builds on the covariance estimators of Hayashi and Yoshida (2005) and Corsi (2006) for the aspects concerning the asynchrony of the data and on the consistent realized variance estimator of Zhang, Mykland and Aït-Sahalia (2005) for the part dealing with the microstructure contamination. The pseudo-aggregation is introduced to efficiently merge these two techniques that deal with two distinct issues concerning the covariance estimation. By eliminating those observations that do not carry any informational content, the pseudo-aggregation gives back series of pseudo-aggregate returns with equal number of data points for which it is relatively easy to define the covariance estimator and its asymptotic properties. Monte Carlo results confirm the unbiasedness and consistency of the CRC and highlight the dramatic improvement with respect the competing estimators in terms of mean squared error.

Natural extensions to the CRC framework are to the possible existence of time delayed covariances as well as the presence of serial correlation in the data which would constitute additional sources of bias if not properly accounted for. These and the extension to the multivariate setting are left as an area for future research.

Acknowledgements

I thank Tim Bollerslev, Fulvio Corsi and Xin Huang for the valuable comments and suggestions. I also thank Giampiero M. Gallo for providing an office space during the summer months in which the manuscript was being prepared.

References


Figure 1: BIAS
Figure 2: log-MSE
Figure 3: BIAS

300 sec, avg. COV

150 sec, avg. COV

60 sec, avg. COV

30 sec, avg. COV

15 sec, avg. COV

5 sec, avg. COV

2 sec, avg. COV

1 sec, avg. COV
Figure 4: log-MSE

300sec, avg COV

150sec, avg COV

60sec, avg COV

30sec, avg COV

15sec, avg COV

5sec, avg COV

2sec, avg COV

1sec, avg COV
Figure 5: BIAS
Figure 6: log-MSE

0.0004NS, avg. COV

0.0008NS, avg. COV

0.0016NS, avg. COV

0.0032NS, avg. COV

0.0084NS, avg. COV
Figure 7: % MSE