A simple and robust measure of microstructure noise from high-frequency financial data

Z. Merrick Li*

University of Amsterdam and Tinbergen Institute

Preliminary version, please do not distribute

Abstract

This paper introduces a new measure of microstructure noise in high-frequency financial data. It takes the form of realized moments of disjoint increments (ReMeDI) of observed log-prices. In particular, we obtain consistent estimators of the second moments of noise under settings that allow for general dependence features. We also derive robust confidence intervals of the ReMeDI estimators. Apart from being consistent, the ReMeDI estimators work remarkably well in finite samples with varying data frequencies, as we demonstrated in our extensive simulation studies. Finally, we present some interesting empirical properties of microstructure noise.

1 Introduction

In the past decade or so, technology has spurred an exponential growth of high-frequency trading operating in time scales from minutes to microseconds. It provides researchers and practitioners abundant data to study the statistical properties of financial returns. One challenging issue, however, is provided by the fact that asset prices embed noise that is induced by various market microstructure frictions, such as information asymmetry and bid-ask spread.

A cornerstone in high-frequency financial econometrics is that the standardized realized moments of observed noisy returns identify the moments of noise when the number

*Email: z.li3@uva.nl. I am indebted to my advisors at University of Amsterdam: Prof. H. Peter Boswijk, Prof. Roger J.A. Laeven and Prof. Michel H. Vellekoop for their help and encouragement. I also benefit from discussions with Yacine Aït-Sahalia and Jean Jacod.
of observations increases within a fixed time interval. For example, the standardized realized volatility (or realized variance) is a consistent estimator of the variance of i.i.d. noise, see Bandi and Russell (2008) and Zhang et al. (2005), see also Li et al. (2016).

The moments of noise play an essential role in estimating the return volatility. In particular, the second moments of noise appear as an asymptotic bias that needs to be corrected to get consistent estimators of the integrated volatility in many de-noise methods, e.g., the two scale realize volatility proposed by Zhang et al. (2005) and the pre-averaging method studied by Jacod et al. (2009), see also Li et al. (2016) and Jacod et al. (2015). The second and fourth moments of noise are the key factors to apply the mean-squared-error optimal sampling theory to estimate the realized volatility in a finite sample, see Aït-Sahalia et al. (2005) and Bandi and Russell (2008).

Two main challenges are raised by the estimation of the moments of noise. First, the noise is usually specified as an i.i.d. process. While the i.i.d. noise captures some stylized properties of observed returns, e.g., the negative first-order autocorrelation, it ignores other dependence structures induced by the trading mechanism such as the clustering of order flows. Second, the magnitude of microstructure noise is very small and the finite sample bias in estimating the moments of noise can be very substantial. Li et al. (2016) show that without bias correction, one would conclude that the microstructure noise has some statistical properties that are hard to explain. The bias term usually involves the (unknown) parameters of the efficient price process. Thus in practice the correction requires some estimates of such parameters, and it inevitably brings more uncertainty.

In this paper, we allow for general dependence features in the microstructure noise. We propose a new class of estimators for the moments of noise. The estimators are formed by taking the realized moments of disjoint increments (ReMeDI) of the observed log-prices. The consistency is achieved by taking relatively large intervals so that only the targeted moments of noise remain and others become negligible asymptotically. The estimators work remarkably well in finite samples. To see the refinement in finite samples, we consider the estimation of the second moments of noise when the efficient price process is a martingale: the efficient price process contributes to the ReMeDI estimator uncorrelated returns since they are disjoint martingale differences, and what remains in the ReMeDI estimator (after taking expectation) is related to the targeted moments of noise.

Empirically, we demonstrate several applications of our ReMeDI estimators using a sample of INTC transaction data. We find that microstructure noise tends to be positively autocorrelated in transaction tick prices. Next, we study the shrinking noise hypothesis. We show that when the sampling frequency increases, the variance of noise estimated by the standardized realized volatility shrinks. However, our ReMeDI estimators return relatively stable estimates — it is the finite sample bias that induces the shrinking effects in the tra-
ditional realized volatility estimator. Finally, we investigate whether a given high-frequency data sample can be considered as noise free by testing whether the variance of noise is zero. We find that data samples with average sampling interval larger than 1.2 seconds are considered as noise-free, and those data samples correspond to the flat part of the volatility signature plots.

The remaining part of the paper is organized as follows. Section 2 discusses the basic models. Section 3 introduces the ReMeDI estimator and presents its theoretical properties. We further demonstrate the performance of our ReMeDI estimators and compare ReMeDI with other estimators via simulation in Section 4. Section 5 presents several interesting empirical applications and Section 6 concludes the paper.

2 Framework and Assumptions

The efficient log-price process \( X \) is a general \( E \)-valued semimartingale, where \( E \) is a Polish space endowed with its Borel \( \sigma \)-field \( \mathcal{E} \), and \( X \) is defined on a filtered probability space \( (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, P) \).

All observations are collected in a fixed time span and, without loss of generality, let it be \([0, 1]\). At stage \( n \), we denote the sampling times by \( 0 = t^n_0 < t^n_1 < ... < t^n_n = 1 \). For any process \( V \), we denote \( V^n_i = V_{t^n_i} \).

In practice, the efficient price process \( X \) is not directly observable; the observed prices embed market microstructure distortions such as bid-ask spreads, information asymmetry, etc. We consider the following general settings that nest several popular models of market microstructure noise, e.g., i.i.d. noise, \( q \)-dependent noise (see Hansen and Lunde (2006), Hautsch and Podolskij (2013)), ARMA(\( p, q \)) noise (see Barndorff-Nielsen et al. (2008)).

Assumption 2.1 (Market Microstructure Noise). The noise process \( \{\varepsilon_i\}_{i \in \mathbb{N}} \) satisfies the following assumptions:

1. \( \mathbb{E}(\varepsilon_i) = 0 \) \( \forall i \in \mathbb{N} \);
2. The noise process \( \{\varepsilon_i\}_{i \in \mathbb{N}} \) is stationary and strongly mixing with mixing coefficients \( |\alpha_k|_{k=0}^{\infty} \) and \( \alpha_k \downarrow 0 \) as \( k \to \infty \).

Remark 2.1. We do not assume any orthogonality conditions between the efficient price and noise process.

---

\(^1\)The mixing coefficients is a sequence satisfying

\[ |\mathbb{P}(A_i \cap A_{i+k}) - \mathbb{P}(A_i)\mathbb{P}(A_{i+k})| \leq \alpha_k \]

for all \( A_i \in \sigma(\varepsilon_j : j \leq i), A_{i+k} \in \sigma(\varepsilon_j : j \geq i + k) \), where \( \sigma(\cdot) \) is the generated \( \sigma \)-algebra.
Figure 1: Illustration of the ReMeDI estimators of variance and covariance of noise.

We denote γ(j) = E(εiεi+j), ∀ j ∈ N. For a sample of size n, we denote $\varepsilon_i^n = \varepsilon_i, \forall 0 \leq i \leq n$. The contaminated observed log-price is given by

$$Y_i^n = X_i^n + \varepsilon_i^n.$$  \hspace{1cm} (1)

### 3 New estimators of moments of noise

In this section, we construct a new class of consistent estimators of noise moments. The estimators are formed by taking the realized moments of disjoint increments (ReMeDI) of the observed log prices. For any process $V$, we denote $\Delta_i^n V = V_i^n - V^n_{i+k}$.

The ReMeDI estimator of the second moments of noise is given by

$$\text{ReMeDI}(j)_n := -\frac{1}{n-3j_n-j+1} \sum_{i=j_n}^{n-2j_n-j} \Delta_i^n X \Delta_i^{n,2j_n} Y, \quad j = 0, 1, 2, \ldots$$ \hspace{1cm} (2)

Figure 1 illustrates the intuition. The observed log returns have two components: the efficient returns and the microstructure noise. The tuning parameter $j_n$ is relatively large so that only $\varepsilon_i^n \varepsilon_i^{n+j}$ remains among the four cross products in $\Delta_i^n X \varepsilon \Delta_i^{n,2j_n}$, i.e., $\Delta_i^n X \varepsilon \Delta_i^{n,2j_n} \varepsilon \approx -\varepsilon_i^n \varepsilon_i^{n+j}$; $j_n$ is still relatively smaller than $n$ so that noise moments dominate the efficient returns, that is, $\Delta_i^n X \varepsilon \Delta_i^{n,2j_n} \varepsilon \approx \Delta_i^n X \varepsilon \Delta_i^{n,2j_n} \varepsilon$. Hence for carefully selected $j_n$, we have $\Delta_i^n X \varepsilon \Delta_i^{n,2j_n} \varepsilon \approx -\varepsilon_i^n \varepsilon_i^{n+j}$.

We can also construct ReMeDI estimators of other moments. The following is a ReMeDI estimator of the fourth moments of noise $E(\varepsilon_0 \varepsilon_j \varepsilon_{j+p} \varepsilon_{j+p+q})$:

$$\text{ReMeDI}(j, p, q)_n := \frac{\sum_{i=j_n}^{n-3j_n-j-p-q} \Delta_i^n X \Delta_i^{n,j} Y \Delta_i^{n,2j} Y \Delta_i^{n,3j} Y}{n-4j_n-j-p-q+1}, \quad j, p, q \in \mathbb{N}.$$ \hspace{1cm} (3)
3.1 Asymptotic theory

**Theorem 3.1.** Let the noise process $\varepsilon$ satisfy Assumption 2.1 and $j_n$ satisfies

$$j_n \to \infty, \quad \frac{n}{j_n} \to \infty. \quad \text{(4)}$$

Assume that there is $\delta > 0$ such that $\mathbb{E}(|\varepsilon|^{8+\delta}) < \infty$. We have the following consistency results:

$$\text{ReMeDI}(j_n) \xrightarrow{P} \mathbb{E}(\varepsilon_0 \varepsilon_j); \quad \text{(5)}$$

$$\text{ReMeDI}(j_n, p, q) \xrightarrow{P} \mathbb{E}(\varepsilon_0 \varepsilon_j \varepsilon_{j+p} \varepsilon_{j+p+q}). \quad \text{(6)}$$

**Proof.** See Appendix A.

**Theorem 3.2** (Central Limit Theorem). Let the noise process $\varepsilon$ satisfy Assumption 2.1 and $j_n, \{\alpha_k\}_{k \geq 1}$ satisfy

$$j_n \to \infty, \quad \frac{n}{j_n} \to \infty, \quad \sum_{k=1}^{\infty} \alpha_k < \infty. \quad \text{(7)}$$

We also assume the moment condition that $\mathbb{E}(|\varepsilon|^{8+\delta}) < \infty$ for some $\delta > 0$. Let

$$\Omega_j := \mathbb{E}\left( (\varepsilon_0 \varepsilon_j - \gamma(j))^2 \right) + 3 (\text{Var}(\varepsilon))^2 + 2 \sum_{i=1}^{\infty} \mathbb{E}(\varepsilon_0 \varepsilon_j - \gamma(j))(\varepsilon_i \varepsilon_{i+j} - \gamma(j)) + 3\gamma(i)^2; \quad \text{(8)}$$

and for $i_n$ satisfies

$$\sum_{k=i_n-3j_n}^{\infty} \alpha_k \to 0, \quad \text{(9)}$$

let

$$\hat{\Omega}_j^n := \frac{1}{n-3j_n} \sum_{i=j_n}^{n-2j_n} \left( \tilde{Y}_{i,i}^{n,j_n} \right)^2 + 2 \sum_{k=1}^{i_n} \tilde{Y}_{i,i}^{n,j_n} \tilde{Y}_{i,i+k}^{n,j_n}; \quad \text{(10)}$$

where $\tilde{Y}_{i,i}^{n,j_n} = -\Delta_{i-j_n}^{n,j_n} Y \Delta_{i+j}^{n,2j_n} Y - \text{ReMeDI}(j_n)$. Then we have the following convergence in distribution:

$$\sqrt{n}(\text{ReMeDI}(j_n) - \gamma(j)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Omega_j); \quad \text{(11)}$$

$$\sqrt{n}(\text{ReMeDI}(j_n) - \gamma(j)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \quad \text{(12)}$$

**Proof.** See Appendix B.
**Remark 3.1.** An alternative estimator of the asymptotic variance $\Omega_j$ is given by

$$
\Omega_j^n := \text{ReMeDI}(0, j, 0)_n + 2 \sum_{i=1}^{n} \left( \text{ReMeDI}(\min\{i, j\}, |i - j|, \min\{i, j\})_n + 3 (\text{ReMeDI}(i)_n)^2 \right) \\
- (2i_n + 1) (\text{ReMeDI}(j)_n)^2.
$$

Indeed, $\Omega_j^n$ is an excellent estimator of $\Omega_j$. However, $\Omega_j$ itself provides poor representation of the finite sample variance of our ReMeDI estimators: it fails to take into account the variance of the efficient price process, which although is asymptotically negligible. Our estimator $\hat{\Omega}_j^n$, however, can capture such finite sample variation.

### 3.2 Finite sample analysis

Now we assume the efficient price is a martingale that is independent of the noise process. Then we can show

$$
E(\text{ReMeDI}(j)_n) = \gamma(j) + K\alpha_{jn},
$$

where $K > 0$ is a constant. Since $\alpha_{jn} \to 0$, the ReMeDI estimators are asymptotically unbiased.

We will further study the finite sample properties of our ReMeDI estimators and compare with the realized volatility (see Li et al. (2016)) and the local averaging estimators (see Jacod et al. (2016)) in the simulation section.

### 4 Simulation studies

#### 4.1 Simulation design

We consider the following general settings for the efficient log-price $X$:

$$
\begin{align*}
\text{d}X_t &= \kappa_1 (\mu_1 - X_t) \text{d}t + \sigma_t \text{d}W_{1,t} + \xi_{1,t} \text{d}N_t; \\
\text{d}\sigma_t^2 &= \kappa_2 (\mu_2 - \sigma_t^2) \text{d}t + \gamma \sigma_t \text{d}W_{2,t} + \xi_{2,t} \text{d}N_t; \\
\text{Corr}(W_{1,t}, W_{2,t}) &= \varrho; \\
\xi_{1,t} &\sim \mathcal{N}(0, \mu_2/10); \ N_t \sim \text{Poi}(\lambda); \ \xi_{2,t} \sim \text{Exp}(\delta).
\end{align*}
$$

(13)
The jumps settings are motivated by empirical facts that jumps in price levels and volatility tend to occur together, see Todorov and Tauchen (2011). We set
\[ \kappa_1 = 0.5; \mu_1 = 1.6; \kappa_2 = 5/252; \mu_2 = 0.04/252; \gamma = 0.05/252; \rho = -0.5; \lambda_1 = 5; \delta = \gamma. \]

We assume an AR(1) noise process, as studied in Aït-Sahalia et al. (2011):
\[ \epsilon_t = V_t + U_t, \quad (14) \]
where \( V \) is centered i.i.d. and \( U \) is an AR(1) process with first order coefficient \( \rho, |\rho| < 1 \). \( V \) and \( U \) are statistically independent. We set
\[ \text{Var}(V) = 2.9 \times 10^{-8}; \text{Var}(U) = 4.3 \times 10^{-8}; \rho = 0.7. \]
Those estimates are borrowed from Aït-Sahalia et al. (2011), except the sign of \( \rho \). We will let \( \rho \) vary in the following simulation studies.

### 4.2 Asymptotic properties

#### 4.2.1 Consistency

In Figure 2, we plot The ReMeDI estimates of the autocovariances of noise based a single simulation path. We consider different data frequencies and model specifications. It is clear from the top panel of Figure 2 that the ReMeDI estimators work well with data at 1s scale across different models. The plots in bottom panel clearly support the consistency theory.

#### 4.2.2 Central limit theorem

Figure 3 presents the QQ plots of the ReMeDI estimators based on our limit distribution theory for the second moments of noise when the efficient price process exhibits jumps in both the price level and volatility. Thanks to the robust estimator of the limiting variance (recall Remark 3.1), we find that our limit distribution represents well the finite sample properties of the ReMeDI estimators in different sampling schemes.

### 4.3 Finite sample properties

In this section, we study the finite sample properties of the ReMeDI estimators. We will compare to the two-step, or bias corrected realized volatility (BCRV) and the bias corrected

\[ ^2 \text{Noise tends to be positively correlated, see Jacod et al. (2016) and Li et al. (2016).} \]
Figure 2: ReMeDI estimates of autocovariances of noise based on a single sample path. The red stars are the true values. In the top panel, the simulated date frequency is $\Delta = 1 \text{s}$ while in the bottom panel $\Delta = 0.1 \text{s}$. From left to right, the four columns correspond to the models with no jumps in neither prices nor volatility, only jumps in prices, only jumps in volatility and with jumps in both prices and volatility.
Figure 3: QQ-plots of ReMeDI estimates of variance and autocovariances of noise based on 1000 simulations, $\Delta_n = s_i = j_i = i_0 = 15$. 
local averaging (BCLA) estimator of the second moments of noise.\footnote{Li et al. (2016) introduce the two-step estimators of the second moments of noise and the integrated volatility. To get the two-step, or bias corrected estimators of the second moments of noise, we first estimate the integrated volatility using the pre-averaging method, however, without taking into account the dependence of noise. Then we use this estimate of integrated volatility to correct the bias in estimating the second moments of noise using realized volatility, see Section 5 in Li et al. (2016) for details. The bias corrected local averaging estimators are obtained in the same way.} The data frequency is fixed at $\Delta = 1\text{s}$.

### 4.3.1 Continuous price and volatility

We consider continuous price and volatility processes, hence $\lambda = 0$. For this class of models, the finite sample theory of estimating the second moments of noise has been developed in Li et al. (2016).

In the left panel of Figure 4, we plot the means of the ReMeDI estimators of noise autocovariances and autocorrelations based on 1000 simulations. We also plot the 95% simulated confidence intervals. On the right panel, we perform the bias corrected realize volatility estimation (BCRV). The two estimators are comparable. The ReMeDI estimator has the advantage of smaller variation for the first several autocovariances, and these are the key statistics to make inference on the integrated volatility.

### 4.3.2 Jumps in prices and volatility

In Figure 5, Figure 6 and Figure 7, we replicate the estimations but allowing for jumps in either price level or volatility, or both. Our ReMeDI estimators are quite robust with slightly larger confidence intervals. But the BCRV and BCLA estimators become much less accurate.

### 4.4 ReMeDI, Realized Volatility and the choice of $j_n$

In this section, we estimate the variance of noise using the standardized realized volatility (RV) and our ReMeDI estimators with varying $j_n$\footnote{Li et al. (2016) introduce the two-step estimators of the second moments of noise and the integrated volatility. To get the two-step, or bias corrected estimators of the second moments of noise, we first estimate the integrated volatility using the pre-averaging method, however, without taking into account the dependence of noise. Then we use this estimate of integrated volatility to correct the bias in estimating the second moments of noise using realized volatility, see Section 5 in Li et al. (2016) for details. The bias corrected local averaging estimators are obtained in the same way.}. The aim of this study is twofold. First, we show the comparative advantages of ReMeDI in estimating the noise variances with different sampling frequencies. Second, we provide some guidance on the choices of $j_n$ to optimize the finite sample performance of our ReMeDI estimators.

Table 1 and Table 2 report the results when the underlying noise process is i.i.d. and AR(1) with coefficient $\rho = 0.7$. The ReMeDI estimator performs well (with small mean-squared-errors) when the data frequency is high, and when the noise deviates from being an i.i.d. process. Moreover, it is relatively robust to the choices of $j_n$. When the data frequencies are lower, the optimal choices of $j_n$ also drops. As a rule of thumb, for example,
Figure 4: Estimation of autocovariances and autocorrelations of microstructure noise by the realized moments of disjoint increments (ReMeDI), bias corrected realized volatility (BSRV) and bias corrected local averaging (BCLA) estimators. The stars are true values. The valid line is the mean estimations based on 1000 simulations. The dashed lines are the simulated 95% confidence intervals. Jumps are absent from both the efficient price process and the volatility process.
Figure 5: Estimation of autocovariances and autocorrelations of microstructure noise by the realized moments of disjoint increments (ReMeDI), bias corrected realized volatility (BSRV) and bias corrected local averaging (BSLA) estimators. The stars are true values. The valid line is the mean estimations based on 1000 simulations. The dashed lines are the simulated 95% confidence intervals. The efficient price process exhibits jumps.
Figure 6: Estimation of autocovariances and autocorrelations of microstructure noise by the realized moments of disjoint increments (ReMeDI), bias corrected realized volatility (BSRV) and bias corrected local averaging (BSLA) estimators. The stars are true values. The valid line is the mean estimations based on 1000 simulations. The dashed lines are the simulated 95% confidence intervals. The volatility process exhibits jumps.
Figure 7: Estimation of autocovariances and autocorrelations of microstructure noise by the realized moments of disjoint increments (ReMeDI), bias corrected realized volatility (BSRV) and bias corrected local averaging (BCLA) estimators. The stars are true values. The valid line is the mean estimations based on 1000 simulations. The dashed lines are the simulated 95% confidence intervals. Both the efficient price process and volatility process exhibit jumps.
a favorite choice of $j_n \approx 10$ when we work on samples with average $\Delta_n \in (0.1 s, 1 s)$, which is typically encountered in tick data.

5 Empirical Studies

5.1 Is microstructure noise i.i.d.?  
We estimate the first 30 autocovariances of noise using the transaction data of INTC obtained from the Trade and Quote (TAQ) database. We perform the estimation on each trading day. Figure 8 shows the patterns of the estimated autocovariances. 

It is clear that microstructure noise tends to be positively autocorrelated. The positive autocorrelations may reflect the clustering of transactions in high-frequency trading. Such patterns are consistent with the findings in Li et al. (2016). However, we obtain the estimates without any bias correction that requires an estimate of the integrated volatility of the efficient returns.

5.2 Magnitude of noise and sampling frequencies: is noise shrinking at higher frequencies?  
A model that relates the sampling frequencies and magnitude of noise is to add a multiplicative shrinking factor $u_n > 0$ to the "genuine noise process" $\varepsilon^n_{i}$ with $u_n \rightarrow 0$ as $n \rightarrow \infty$, see Chapter 7 in Aït-Sahalia and Jacod (2014) and Chapter 16 in Jacod and Protter (2011). Empirically, to the best of our knowledge, the shrinking effect is neither confirmed nor rejected except some heuristic evidence supporting shrinking noise in Kalnina and Linton (2008).

Motivated by the finite sample analysis in Li et al. (2016), we show that the shrinking magnitudes of microstructure noise might be a consequence of finite sample bias. In Figure 9, we plot the estimates of noise variance using both the standardized realized volatility (RV) and ReMeDI estimators for INTC on February 2, 2016 in subsamples with varying frequencies.

We observe in Figure 9 that the RV estimates increase as the sampling frequency decreases — this is a supportive evidence of shrinking noise — the magnitude of noise drops as the data frequency becomes high. But according to the finite sample analysis by Li et al. (2016), the upward slope is most likely induced by finite sample bias, when (a fraction of) integrated volatility starts to bite the moments of noise. Interestingly, the ReMeDI estimates are relatively stable with wide confidence band at lower frequency samples. Moreover, most of the RV estimates fall out of the confidence intervals of ReMeDI estimator. Hence our ReMeDI estimates tend to reject the shrinking noise hypothesis.
<table>
<thead>
<tr>
<th></th>
<th>$\Delta n = 0.1\text{s}$</th>
<th></th>
<th>$\Delta n = 1.0\text{s}$</th>
<th>$\Delta n = 10.0\text{s}$</th>
<th>$\Delta n = 30.0\text{s}$</th>
<th>$\Delta n = 60.0\text{s}$</th>
<th>$\Delta n = 300.0\text{s}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
</tr>
<tr>
<td></td>
<td>$j_{n1} = 10$</td>
<td>$j_{n2} = 16$</td>
<td>$j_{n3} = 20$</td>
<td>$j_{n1} = 10$</td>
<td>$j_{n2} = 16$</td>
<td>$j_{n3} = 20$</td>
<td>$j_{n1} = 10$</td>
</tr>
<tr>
<td>Bias</td>
<td>1.58e-09</td>
<td>-1.17e-11</td>
<td>1.58e-11</td>
<td>2.59e-11</td>
<td>1.60e-08</td>
<td>-4.08e-10</td>
<td>9.37e-10</td>
</tr>
<tr>
<td>S.t.d.</td>
<td>8.44e-10</td>
<td>4.92e-10</td>
<td>6.94e-10</td>
<td>9.24e-10</td>
<td>8.58e-09</td>
<td>9.07e-09</td>
<td>1.71e-08</td>
</tr>
<tr>
<td>MSE.</td>
<td>3.21e-18</td>
<td>2.42e-19</td>
<td>4.82e-19</td>
<td>8.55e-19</td>
<td>3.31e-16</td>
<td>8.23e-17</td>
<td>2.94e-16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\Delta n = 10.0\text{s}$</td>
<td></td>
<td>$\Delta n = 30.0\text{s}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
</tr>
<tr>
<td></td>
<td>$j_{n1} = 4$</td>
<td>$j_{n2} = 10$</td>
<td>$j_{n3} = 16$</td>
<td>$j_{n1} = 4$</td>
<td>$j_{n2} = 10$</td>
<td>$j_{n3} = 16$</td>
<td>$j_{n1} = 4$</td>
</tr>
<tr>
<td>Bias</td>
<td>1.51e-07</td>
<td>2.70e-09</td>
<td>2.75e-10</td>
<td>1.37e-08</td>
<td>4.59e-07</td>
<td>3.32e-08</td>
<td>2.43e-08</td>
</tr>
<tr>
<td>S.t.d.</td>
<td>7.56e-08</td>
<td>6.58e-08</td>
<td>2.83e-07</td>
<td>5.49e-07</td>
<td>2.58e-07</td>
<td>3.58e-07</td>
<td>1.35e-06</td>
</tr>
<tr>
<td>MSE.</td>
<td>2.86e-14</td>
<td>4.33e-15</td>
<td>8.01e-14</td>
<td>3.01e-13</td>
<td>2.77e-13</td>
<td>1.29e-13</td>
<td>1.83e-12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\Delta n = 60.0\text{s}$</td>
<td></td>
<td>$\Delta n = 300.0\text{s}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
</tr>
<tr>
<td></td>
<td>$j_{n1} = 2$</td>
<td>$j_{n2} = 4$</td>
<td>$j_{n3} = 6$</td>
<td>$j_{n1} = 2$</td>
<td>$j_{n2} = 4$</td>
<td>$j_{n3} = 6$</td>
<td>$j_{n1} = 2$</td>
</tr>
<tr>
<td>Bias</td>
<td>9.60e-07</td>
<td>2.41e-09</td>
<td>2.57e-08</td>
<td>1.99e-07</td>
<td>4.75e-06</td>
<td>4.91e-07</td>
<td>7.32e-07</td>
</tr>
<tr>
<td>S.t.d.</td>
<td>5.23e-07</td>
<td>3.65e-07</td>
<td>1.04e-06</td>
<td>1.91e-06</td>
<td>2.47e-06</td>
<td>4.25e-06</td>
<td>1.10e-05</td>
</tr>
<tr>
<td>MSE.</td>
<td>1.20e-12</td>
<td>1.33e-13</td>
<td>1.09e-12</td>
<td>3.68e-12</td>
<td>2.87e-11</td>
<td>1.83e-11</td>
<td>1.21e-10</td>
</tr>
</tbody>
</table>

Table 1: ReMeDI and Realized Volatility (RV) estimation of variance of i.i.d. noise.
<table>
<thead>
<tr>
<th></th>
<th>$\Delta_n = 0.1s$</th>
<th></th>
<th>$\Delta_n = 1.0s$</th>
<th>$\Delta_n = 10.0s$</th>
<th>$\Delta_n = 30.0s$</th>
<th>$\Delta_n = 60.0s$</th>
<th>$\Delta_n = 300.0s$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
<td>ReMeDI</td>
<td>RV</td>
</tr>
<tr>
<td></td>
<td>$j_{n1} = 10$</td>
<td>$j_{n2} = 16$</td>
<td>$j_{n3} = 20$</td>
<td>$j_{n1} = 10$</td>
<td>$j_{n2} = 16$</td>
<td>$j_{n3} = 20$</td>
<td>$j_{n1} = 10$</td>
</tr>
<tr>
<td>Bias</td>
<td>-2.85e-08</td>
<td>-1.41e-09</td>
<td>-1.35e-10</td>
<td>-9.40e-11</td>
<td>-1.46e-08</td>
<td>-1.14e-09</td>
<td>-9.82e-10</td>
</tr>
<tr>
<td>S.t.d.</td>
<td>8.51e-10</td>
<td>6.07e-10</td>
<td>8.40e-10</td>
<td>1.05e-09</td>
<td>8.11e-09</td>
<td>9.37e-09</td>
<td>1.82e-08</td>
</tr>
<tr>
<td>MSE.</td>
<td>8.15e-16</td>
<td>2.35e-18</td>
<td>7.23e-19</td>
<td>1.11e-18</td>
<td>2.78e-16</td>
<td>8.90e-17</td>
<td>3.33e-16</td>
</tr>
<tr>
<td></td>
<td>$\Delta_n = 60.0s$</td>
<td></td>
<td>$\Delta_n = 300.0s$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>1.22e-07</td>
<td>-1.16e-08</td>
<td>6.41e-09</td>
<td>5.38e-08</td>
<td>4.26e-07</td>
<td>7.57e-09</td>
<td>8.86e-08</td>
</tr>
<tr>
<td>S.t.d.</td>
<td>8.29e-08</td>
<td>6.61e-08</td>
<td>2.84e-07</td>
<td>5.17e-07</td>
<td>2.36e-07</td>
<td>3.42e-07</td>
<td>1.46e-06</td>
</tr>
<tr>
<td>MSE.</td>
<td>2.18e-14</td>
<td>4.50e-15</td>
<td>8.06e-14</td>
<td>2.70e-13</td>
<td>2.37e-13</td>
<td>1.17e-13</td>
<td>2.15e-12</td>
</tr>
</tbody>
</table>

Table 2: ReMeDI and Realized Volatility (RV) estimation of variance of AR(1) noise. AR(1) coefficient $\rho = 0.7$. 

17
Figure 8: ReMeDI estimates of autocovariances of microstructure noise for INTC transaction prices in February 2016. The dash lines are the 95% confidence intervals. We set $j_n = 6, i_n = 20$. 
Figure 9: ReMeDI and Realized Volatility (RV) estimates of variance of noise for INTC on February 2, 2016. We subsample the data at every $k$ observations with $k = 1$ (corresponding to the full sample), 3, 10, 30, 100, 200, 300, 500, 1000; the average inter-observations time lags $\Delta_n$ are 0.1s, 0.4s, 1.2s, 3.5s, 11.7s, 23.5s, 35.3s, 58.8s, 117.6s, respectively.
5.3 At what frequencies is the sample free of noise?

The accumulation of microstructure noise at high-frequency data is directly supported by the volatility signature plots, in which, the realized volatility diverges as the inter-observations time $\Delta_n$ converges to zero, see, e.g., Bandi and Russell (2008), Hansen and Lunde (2006).

The bottom panel of Figure 10 contains a volatility signature plot for INTC on February 2, 2016. It is evident that the plot is upward sloping at higher frequencies. However, it is not trivial to decide at what sampling frequencies the sample is considered as noise free so that the volatility of stock returns can be measured by the simple, model-free realized volatility estimator. Or, equivalently put, we need a decisive rule to determine at what frequencies the noise needs to be corrected by a de-noise method.

The top panel of Figure 10 depicts a concrete procedure: we construct the confidence intervals for the ReMeDI estimate of noise variance and then decide whether it is significantly different from zero. We first observe that the confidence band becomes wider when data frequencies are lower. For samples with inter-observation lag $\Delta_n$ greater than 1.2s (corresponding to the first red circle), it is relatively safe to conclude that the volatility of the efficient price dominates the noise component hence the samples can be considered as “free” of noise. Turning to the volatility signature plot, those samples correspond to the flatter part (blue nodes connected by dashed line) of the plot.

6 Conclusion

Financial returns embed microstructure noise that accumulates at high frequencies. Studying the statistical properties of microstructure noise is of great importance in estimating the financial return variability. In this paper, we introduce a new class of estimators of moments of noise under general specification settings. Our new estimators have nice theoretical properties, and are robust to different sampling schemes in practice.

In this paper, we try to balance the theoretical and empirical results to provide suggestions for future research. Our empirical study recovers some interesting properties of microstructure noise in the examined stock. However, an abundant study of the empirical properties of microstructure noise is still pending. Our empirical study reveals that the popular de-noise method that hinges on i.i.d. noise needs revision, and this is partially done in a companion paper Li (2017), see also Li et al. (2016).
Figure 10: ReMeDI estimates of variance of noise and volatility signature plot for INTC on February 2, 2016. We subsample the data at every $k$ observations with $k = 1$ (corresponding to the full sample), 3, 10, 30, 100, 200, 300, 500, 1000; the average inter-observations time lags $\Delta_n$ are 0.1 s, 0.4 s, 1.2 s, 3.5 s, 11.7 s, 23.5 s, 35.3 s, 58.8 s, 117.6 s, respectively.
Appendix: Technical Proofs

A Proof of Proposition 3.1

Proof. 1. We first show (5). Let

\[ c_i = \begin{cases} 
X_{i+j}^n - X_{i}^n & j_n \leq i \leq n - 2j_n - j \\
0 & \text{otherwise.} 
\end{cases} \]

Then

\[ \sum_{i=j_n}^{n-j_n-j} (X_i^n - X_{i+j_n}^n)(\varepsilon_{i+j+n}^n - \varepsilon_{i}^n) = \sum_{i=j_n}^{n-j_n-j} (c_i - c_{i+j_n})\varepsilon_i^n. \]

Conditional on the efficient price process, we have

\[ \mathbb{E} \left( \left( \sum_{i=j_n}^{n-j_n-j} (c_i - c_{i+j_n})\varepsilon_i^n \right)^2 \bigg| X \right) \leq \sum_{i=j_n}^{n-j_n-j} (c_i - c_{i+j_n})^2 \text{Var} (\varepsilon) \\
+ 2 \sum_{i=j_n}^{n-j_n-j} \sum_{k>i}^{n-j_n-j} |(c_i - c_{i+j_n})(c_k - c_{k+j_n})| \text{Cov}(\varepsilon_i^n, \varepsilon_k^n) \]

\[ \leq \sum_{i=j_n}^{n-j_n-j} (c_i - c_{i+j_n})^2 \left( \text{Var} (\varepsilon) + 2 \sum_{h=1}^{n-2j_n-j} \alpha_h \right) \\
\leq 2j_n[X, X] \left( \text{Var} (\varepsilon) + 2 \sum_{h=1}^{n-2j_n-j} \alpha_h \right). \]

The second inequality follows from the Cauchy-Schwartz inequality. The assumption (4) implies the above is of order \( O_p(j_n). \) The term \( \sum_{i=j_n}^{n-j_n-j} (X_i^n - X_{i+j_n}^n)(X_{i+j+n}^n - X_{i+j}^n) \) is bounded by \( j_n[X, X] \) by, again, Cauchy-Schwartz inequality.

Now we will find the estimates of

\[ \sum_{i=j_n}^{n-j_n-j} (\varepsilon_i^n - \varepsilon_{i+j_n}^n)(\varepsilon_{i+j+n}^n - \varepsilon_{i+j}^n). \]

Denote

\[ \mu_{j, j_n} := \mathbb{E} \left( (\varepsilon_i^n - \varepsilon_{i+j_n}^n)(\varepsilon_{i+j+n}^n - \varepsilon_{i+j}^n) \right) = -\gamma(j) + 2\gamma(j + j_n) - \gamma(j + 2j_n). \]

The stationary assumption implies \( \mu_{j, j_n} \) does not depend on \( i. \) For some large \( h, \) we
have

\[
\begin{align*}
&\left|\text{Cov}\left(\varepsilon_i^n - \varepsilon_{i-j}^n, \varepsilon_{i+j}^n - \varepsilon_{i+j}^n, \varepsilon_{i-h}^n - \varepsilon_{i-h}^n, \varepsilon_{i+h+j}^n - \varepsilon_{i+h+j}^n\right)\right| \\
&= \left|\mathbb{E}\left[\left(\varepsilon_i^n - \varepsilon_{i-j}^n, \varepsilon_{i+j}^n - \varepsilon_{i-j}^n\right) \cdot \left(\varepsilon_{i+h}^n - \varepsilon_{i+h}^n, \varepsilon_{i+h+j}^n - \varepsilon_{i+h+j}^n\right)\right] - \mu_{j,n} \cdot \left(\varepsilon_{i+h}^n - \varepsilon_{i+h}^n, \varepsilon_{i+h+j}^n - \varepsilon_{i+h+j}^n\right)\right| \\
&\leq \sqrt{\text{Var}\left(\varepsilon_i^n - \varepsilon_{i-j}^n\right) \cdot \text{Var}\left(\varepsilon_{i+j}^n - \varepsilon_{i-j}^n\right) \cdot \text{Var}\left(\varepsilon_{i+h}^n - \varepsilon_{i+h}^n\right) \cdot \text{Var}\left(\varepsilon_{i+h+j}^n - \varepsilon_{i+h+j}^n\right)} \\
&\leq 6\sqrt{\alpha_h \text{Var}\left(\varepsilon_i^n - \varepsilon_{i-j}^n\right) \cdot \text{Var}\left(\varepsilon_{i+j}^n - \varepsilon_{i-j}^n\right)}.
\end{align*}
\]

The second inequality is due to Lemma VIII 3.102 in Jacod and Shiryaev (2003).

\[\mathbb{E}\left(|\varepsilon|^8 + \delta\right) < \infty \text{ implies } \text{Var}\left(\varepsilon_i^n - \varepsilon_{i-j}^n, \varepsilon_{i+j}^n - \varepsilon_{i-j}^n\right) < \infty, \text{ this implies the moment }\]

\[\mathbb{E}\left(\sum_{i=j}^{n-j-j} (\varepsilon_i^n - \varepsilon_{i-j}^n, \varepsilon_{i+j}^n - \varepsilon_{i-j}^n) \cdot \text{Var}\left(\varepsilon_{i+h}^n - \varepsilon_{i+h}^n\right) \cdot \text{Var}\left(\varepsilon_{i+h+j}^n - \varepsilon_{i+h+j}^n\right)\right)^2 \text{ is uniformly bounded.} \]

Therefore \[\frac{1}{n-2j_n+1} \sum_{i=j_n}^{n-j-j} (\varepsilon_i^n - \varepsilon_{i-j}^n, \varepsilon_{i+j}^n - \varepsilon_{i-j}^n) = -\gamma(j) + O_p(1/\sqrt{n-2j_n+1}) + 2\gamma(j + j_n) - \gamma(j) + 2j_n). \text{ Now the proof of (5) is finished.} \]

2. Now we sketch the proof (6) which is quite similar to the proof of (6). The following

\[\sum_{i=j_n}^{n-j-j-p-q} \Delta_{i-j}^{n,j_n} X \Delta_{i+j+p}^{n,2j_n} \Delta_{i+j+p+q}^{n,3j_n} = O_p(j_n)\]

is a consequence of the Cauchy-Schwartz inequality and the classic result for semi-martingale: \[\sum_{i=1}^{\lfloor t/\Delta_t \rfloor} |\Delta_t^X| P \sum_{\Delta=|t|}^{\lfloor |t|/\Delta \rfloor} \|X_i^\Delta\|^p \forall p > 2. \]

We can also show all cross terms are of order \(O_p(j_n)\), using the same techniques as we employed above to show \(\sum_{i=j_n}^{n-j-j} (X_i^n - X_{i-j}^n) (\varepsilon_{i+j}^n - \varepsilon_{i+j}^n) = O_p(j_n)\).

Now we show

\[\frac{1}{n-4j_n-p-j-q+1} \sum_{i=j_n}^{n-j-j-p-q} \Delta_{i-j}^{n,j_n} X \Delta_{i+j+p}^{n,2j_n} \Delta_{i+j+p+q}^{n,3j_n} = O_p(j_n)\]

Among all the 16 cross products, \[\frac{1}{n-4j_n-p-j-q+1} \sum_{i=j_n}^{n-j-j-p-q} \varepsilon_{i+j}^n \text{Var}(\varepsilon_{i+j}^n - \varepsilon_{i+j}^n + \varepsilon_{i+j+p+q}^n) \]

\[\mathbb{E}(\varepsilon_0 e_{i+j+p} e_{i+j+p+q}) \] follows from the moment condition \(\mathbb{E}(|\varepsilon|^8 + \delta) < \infty. \]

For other cross products, the maximal index is at least \(j_n\) larger than the second largest index, for instance, \(\varepsilon_i^n \varepsilon_{i+j}^n \varepsilon_{i+j+p+q}^n \varepsilon_{i+j+p+q+3j_n}^n\). Let the indices be \(k_1 \leq k_2 \leq k_3 \leq k_4\), then we choices of the indices imply \(k_4 - k_3 \geq j_n\). Then our moment condition,
together with Lemma VIII 3.102 in Jacod and Shiryaev (2003) imply
\[
\left| \mathbb{E}\left( e_{k_1}^n e_{k_2}^n e_{k_3}^n e_{k_4}^n \right) \right| \leq K \alpha_{j_n}
\]
for some constant \( K \). And this finishes the proof.

\[ \Box \]

B Proof of Theorem 3.2

Proof. To simplify the notations, we omit the superscript \( n \) in \( \varepsilon^n \). First, we note
\[
\mathbb{E}\left( (\varepsilon_i - \varepsilon_{i-j_n})(\varepsilon_{i+j+2j_n} - \varepsilon_{i+j}) + \gamma(j) \right)^2
\]
\[
= \mathbb{E}\left( (\varepsilon_i \varepsilon_{i+j} - \gamma(j))^2 \right) + \mathbb{E}\left( (\varepsilon_{i-j_n} \varepsilon_{i+j+2j_n} - \varepsilon_{i-j_n} \varepsilon_{i+j})^2 \right)
\]
\[
+ 2 \mathbb{E}\left( (\varepsilon_i \varepsilon_{i+j} - \gamma(j)) (\varepsilon_{i-j_n} \varepsilon_{i+j+2j_n} - \varepsilon_{i-j_n} \varepsilon_{i+j}) \right).
\]
Following Lemma VIII 3.102 in Jacod and Shiryaev (2003), we have
\[
\mathbb{E}\left( (\varepsilon_{i-j_n} \varepsilon_{i+j+2j_n} - \varepsilon_i \varepsilon_{i+j+2j_n} - \varepsilon_{i-j_n} \varepsilon_{i+j})^2 \right)
\]
\[
= \mathbb{E}\left( (\varepsilon_{i-j_n} \varepsilon_{i+j+2j_n})^2 \right) + \mathbb{E}\left( (\varepsilon_i \varepsilon_{i+j+2j_n})^2 \right)
\]
\[
+ 2 \mathbb{E}\left( (\varepsilon_{i-j_n} \varepsilon_{i+j}) (\varepsilon_i \varepsilon_{i+j+2j_n}) \right) - 2 \mathbb{E}\left( (\varepsilon_{i-j_n} \varepsilon_{i+j+2j_n}) (\varepsilon_i \varepsilon_{i+j+2j_n}) \right)
\]
\[
- 2 \mathbb{E}\left( (\varepsilon_{i-j_n} \varepsilon_{i+j+2j_n}) (\varepsilon_{i-j_n} \varepsilon_{i+j}) \right)
\]
\[
= 3 \text{Var}(\varepsilon)^2 + o(1),
\]
and
\[
\mathbb{E}\left( (\varepsilon_i \varepsilon_{i+j} - \gamma(j)) (\varepsilon_{i-j_n} \varepsilon_{i+j+2j_n} - \varepsilon_i \varepsilon_{i+j+2j_n} - \varepsilon_{i-j_n} \varepsilon_{i+j}) \right) = o(1).
\]
Therefore, we have
\[
\mathbb{E}\left( (\varepsilon_i - \varepsilon_{i-j_n})(\varepsilon_{i+j+2j_n} - \varepsilon_{i+j}) + \gamma(j) \right)^2
\]
\[
= \mathbb{E}\left( (\varepsilon_i \varepsilon_{i+j} - \gamma(j))^2 \right) + 3 \text{Var}(\varepsilon)^2 + o(1). \quad (15)
\]
Moreover, for any $k \geq 1$,

\[
E\left(\left(\varepsilon_i - \varepsilon_{i-j_n}\right)\left(\varepsilon_{i+j+2j_n} - \varepsilon_{i+j}\right)^2 + \gamma(j)\right)\left(\left(\varepsilon_{i+k} - \varepsilon_{i+k-j_n}\right)\left(\varepsilon_{i+k+j+2j_n} - \varepsilon_{i+k+j}\right) + \gamma(j)\right)
\]
\[
= E\left(\left(\varepsilon_i \varepsilon_{i+j} - \gamma(j)\right)\left(\varepsilon_{i+k} \varepsilon_{i+k+j} - \gamma(j)\right)\right) + E\left(\varepsilon_{i-j_n} \varepsilon_{i-j_n+j} \varepsilon_{i+k-j_n} \varepsilon_{i+k+j} \varepsilon_{i+k+j+2j_n}\right) + o(1)
\]
\[
= E\left(\left(\varepsilon_i \varepsilon_{i+j} - \gamma(j)\right)\left(\varepsilon_{i+k} \varepsilon_{i+k+j} - \gamma(j)\right)\right) + 3\gamma(k)^2 + o(1).
\]

Since $\sum_{k=1}^{\infty} a_k < \infty$, we see that $\Omega_j$ is well defined, again, by Lemma VIII 3.102 in Jacod and Shiryaev (2003). Combined with the above results, we establish (11). To show $\hat{\Omega}_j \overset{p}{\to} \Omega_j$, we first note our choice of $j_n$ and the convergence in (5) imply

\[
\hat{\Omega}_j^n = \frac{1}{n-3j_n+j+1} \sum_{i-j_n}^{n-j_n} \left(\bar{\varepsilon}_i^{n,j_n}\right)^2 + 2 \sum_{k=1}^{j_n} \left(\bar{\varepsilon}_i^{n,j_n}\right)\left(\bar{\varepsilon}_{i+k}^{n,j_n}\right) + o_p(1),
\]

where $\bar{\varepsilon}_i^{n,j_n} = -\Delta_{i-j_n}^{n,j_n} \Delta_{i+j}^{n,j_n} \varepsilon - \gamma(j)$. (9) implies $\sum_{i-j_n+1}^{\infty} \left(\bar{\varepsilon}_i^{n,j_n}\right)\left(\bar{\varepsilon}_{i+k}^{n,j_n}\right) = o(1)$, now $\hat{\Omega}_j^n \overset{p}{\to} \Omega_j$ follows from the convergence of the fourth sample moments of noise and the moment condition $E\left(|\varepsilon|^{8+\delta}\right) < \infty$. \qed
References


