A Simple Measure of Microstructure Noise

Z. Merrick Li*

University of Amsterdam and Tinbergen Institute

This version: February 27, 2018

Abstract

This paper introduces a simple and intuitive measure of microstructure noise, the deviation of observable transaction prices from fundamental values. We measure the moments of noise, in particular, the variances and autocovariances under a general nonparametric setting. We demonstrate the intrinsic consistency of the proposed estimators without restrictions on data frequencies, and characterize the limit distributions under infill asymptotics. Simulation studies show the robustness of the proposed estimators to data frequencies and model specifications.

The new econometric techniques provide two liquidity measures that gauge the *instantaneous* and *average* bid-ask spread with potentially autocorrelated order flows. While being flexible with the autocorrelation structures, the new estimators only employ the transaction prices thus do not require any knowledge of the order flows. Empirically we find that microstructure noise in transaction data tends to be positively autocorrelated. Such positive autocorrelation induces sharp discrepancies among the bid-ask spread measures: The average measures are persistently larger than the instantaneous ones, whereas the classic Roll measures are further downward biased. Moreover, the intraday spreads have a prominent *L*-shape: The magnitude is much larger at the beginning of the trading day, and it is associated with extremely large transactions.

Keywords: Bid-ask spread, execution costs, finite sample bias, high-frequency data, liquidity, intraday pattern, microstructure noise, mixing sequence, Roll's measure

1 Introduction

Observed asset prices embed frictions induced by diverse microstructure effects such as transaction costs, price discreteness, inventory holdings, information asymmetry, etc. Financial

*Email: z.merrick.li@gmail.com. I am indebted to my advisors at University of Amsterdam: Peter Boswijk, Roger Laeven and Michel Vellekoop for their guidance and encouragement. I benefit from discussions with Yacine Aït-Sahalia, Peter Reinhard Hansen, Jean Jacod, Ilze Kalnina, Oliver Linton and Albert Menkveld, as well as seminar participants at the 2017 North American Summer Meeting and 2017 European Meeting of the Econometric Society, University of Cambridge. I would like to thank Oliver Linton for his hospitality during my visit to University of Cambridge, where part of the paper was developed. The financial support from C.Willems Stichting is gratefully acknowledged.

models decompose the observed prices[1] into a semimartingale and stationary components. The semimartingale component is identified with the implicit *efficient price*; the stationary component, the deviation between the efficient price and the observed transaction prices, is termed the *microstructure noise*.

Two strands of literature deal with the modelling and measuring of the microstructure noise. In *empirical microstructure*, microstructure noise typically resembles the bid-ask spread, which can be further decomposed into components due to order processing cost, adverse information and inventory costs, see Hasbrouck (2007) for an insightful review. There are several classes of statistical models that measure the bid-ask spreads and its components. Roll (1984) method is based on the covariances of returns, see also Choi et al. (1988), Stoll (1989) for extended Roll measures. Regression models employ the order flows (trading direction indicators of buying or selling) and other variables, see Huang and Stoll (1997) and Lin et al. (1995). The Bayesian Gibbs approach is proposed by Hasbrouck (2004, 2009). Chen et al. (2017a,b) develop a semiparametric estimation method of bid-ask spread based on empirical characteristic function. The second strand of literature belongs to the field of *financial econometrics*, in which the inference of the efficient price process, e.g., volatility, jumps, is the primary concern. The economic modelling of microstructure noise is not explicit, but its presence makes the inference of the underlying efficient price more challenging. This leads to the development of several denoise methods, in which the moments of noise and parameters of the efficient price are jointly estimated, e.g., the two/multi-scales realized volatility Zhang et al. (2005), Zhang (2006), Aït-Sahalia et al. (2011); finite sample treatment by Bandi and Russell (2008), Bandi and Russell (2006); maximum likelihood estimator by Aït-Sahalia et al. (2005), Xiu (2010); pre-averaging method developed in Podolskij and Vetter (2009), Jacod et al. (2009); realized kernel by Hansen and Lunde (2006), Barndorff-Nielsen et al. (2008).

The above approaches confront several facts. First, microstructure data are remarkably plentiful. The data richness would favour some flexible and robust methods, making the estimation based on asymptotic approximations appealing. However, most empirical microstructure literature do not fully exploit this data advantage. Second, in financial econometrics, microstructure noise is frequently modelled as an i.i.d. process. Yet trading practices and microeconomic mechanisms generate more complicated microstructure noise. Ignoring the rich structure of microstructure noise makes the estimation results lack economic insights, providing little guidance to investors and financial regulators. The last concern is primarily of a practical nature. Econometric practices with microstructure data should account for sampling schemes and data frequencies. The former may affect price behaviour and many microstructure effects, while the latter imposes limits on the accuracy of asymptotic estimators in a finite sample.

We introduce a general approach to measure microstructure noise in a nonparametric setting. We estimate the moments of noise by the Realized moMents of Disjoint Increments (ReMeDI) of observed transaction prices. The underlying efficient price follows a semimartin-

[1]Price always refers to the logarithmic price unless stated otherwise.

gale that accommodates stochastic volatility, jumps, etc. The microstructure noise is a strongly mixing sequence thus can be serially dependent to capture, for instance, serial autocorrelations induced by clustered order flows. The efficient price and noise could also be correlated to reflect informational effects. *Statistically*, we provide a general method to separate a mixing sequence from a semimartingale; from an *economic* point of view, we identify the components of asset prices arise from market frictions.

We first show, without imposing any restrictions on data frequency that ReMeDI provides consistent estimators of arbitrary second and third moments of microstructure noise when the efficient price is a martingale process and is independent of the noise process. The identification strategy is very intuitive: after taking the realized moments (of the observed prices) over *disjoint* intervals, the efficient price, were it a martingale, constitutes *martingale differences of disjoint intervals* of the ReMeDI estimators. Therefore, the ReMeDI approach effectively "removes" the efficient price thanks to its martingale property, a property that is irrespective of the data frequencies. What remains in the ReMeDI estimators are components of the observed transaction price but the martingale part, which is identified with microstructure noise. ReMeDI further controls and tunes the "length" and "distance" of the intervals to eliminate the redundant moments of microstructure noise. In the end, only the targeted moment remains. Thus the ReMeDI approach follows a simple design and it is invariant to sampling frequencies. It can be used by asset pricers with daily or coarser returns and researchers in market microstructure working on millisecond prices. Next, under *infill asymptotics* when the data frequency increases without bound within a fixed time span, we derive consistent ReMeDI estimators of higher order moments of microstructure noise and obtain the limit distributions of the second moments estimators. The intuition of the identification and estimation under infill asymptotics is different: In the limit, the mesh of observation grid is shrinking to zero, thus the variation of returns over very brief time intervals is largely due to the microstructure effects. As a result, the moments of observed returns can identify the the moments of microstructure noise.

To demonstrate the applications of the proposed estimators, we propose two general liquidity measures to gauge the *instantaneous* and *average* effective bid-ask spread. The measures have several attractive features. First, the measures are model free thus avoid potential misspecification. Second, they are robust to the dynamic patterns in the order flows yet do not require any prior information of the order flows or trading directions. Third, the second liquidity measure is explicitly designed to gauge the average effective spreads, the deviation of the average transaction prices from the fundamental values in a sequence of orders that could stem from a split large order. We develop the corresponding ReMeDI estimators of the two liquidity measures. The ReMeDI approach has several advantages. First, the ReMeDI estimators are easy to implement and computationally very efficient. This is an attractive feature to deal with intraday high-frequency data. Next, the estimators are frequency invariant thus can be applied in a variety of data sets and research areas. Moreover, the estimators enjoy a feasible central limit theorem to assess the accuracy of the estimates using high-frequency data.

To demonstrate the robustness of the ReMeDI approach and make it appealing to distinct

audiences, we conduct extensive simulation studies in several benchmark environments of financial econometrics and market microstructure. We first show that the ReMeDI estimators are robust to extreme events, featured jumps in both the price level and volatilities. The second numerical experiment is motivated by Hasbrouck and Ho (1987). The ReMeDI estimators are able to recover the bid-ask spread in the presence of autocorrelated order flows and price adjustment of midquotes. Of particularly intriguing is a scenario in which a random walk efficient price and an AR(2) bid-ask spread generate returns process that is visually indistinguishable from an MA(1) process, the latter is often perceived as the "empirical support" to postulate i.i.d. spreads. The last numerical study is based on Hendershott et al. (2013) that explicitly model the correlation of the efficient price and noise [2]. The ReMeDI estimators can still identify the noise parameters with great accuracy.

We apply our new liquidity measures to the INTC and KO intraday transaction prices. We find positively autocorrelated microstructure noise in both stocks. The average measure of bid-ask spread is larger than the instantaneous measure. Both measures are persistently larger than the classic Roll measure, which is downward biased in the presence of positively autocorrelated noise. The various measures of spreads exhibit prominent intraday pattern. The measures are significantly larger in the opening of the trading day, and they are accompanied by large orders.

This paper introduces an econometric approach to richer microstructure models. It aims to integrate the (empirical) financial market microstructure and (high-frequency) financial econometrics. It is, however, not the first attempt to push towards the integration of the two fields. Diebold and Strasser (2013) focus on the correlation of efficient price and noise in several leading microstructure models, and study the implication for integrated volatility estimation. Bandi et al. (2017) develop a novel measure of the staleness of stock returns under the infill asymptotic framework. Jacod et al. (2017) propose a class of estimators of the noise moments using high-frequency data.[3] Bollerslev et al. (2018) study the relationship between trading volume and return volatility around important public news announcements using intraday high-frequency data. The study relies critically on the high-frequency econometrics techniques to identify jumps. Da and Xiu (2017) advocate the quasi maximum likelihood approach to estimate both the volatility and the autocovariances of moving-average microstructure noise.

The rest of the paper proceeds as follows. Section 2 introduces a general framework for the microstructure noise and proposes two naive estimators of the autocovariances of noise to motivate the ReMeDI approach. Section 3 presents the ReMeDI estimators and the frequency-free and high-frequency asymptotic theories. Section 4 introduces two model-free measures of the bid-ask spread, and the corresponding ReMeDI estimators of the two measures are developed in Section 5. Section 6 provides extensive simulation examples to examine the finite sample performances of the ReMeDI estimators and Section 7 explains the finite sample robustness. Section 8 contains empirical studies based on the transaction prices of two stocks and provides some interesting empirical properties of the effective spreads. Section 9 concludes

[2] It is termed the *pricing error* in Hendershott et al. (2013).

[3] In the simulation study, we will compare the estimators proposed by Jacod et al. (2017) to the ReMeDI estimators.

the paper. All mathematical proofs are collected in the Appendix.

2 Microstructure Noise and Two Naive Estimators

The basic model considered in this paper is

$$Y = X + \varepsilon, \tag{1}$$

where Y is the observed price of a financial asset, X is the efficient price (or fundamental value) tjat prevails in a frictionless market, ε is the microstructure noise that measures how closely the observed price conforms to the efficient price. Section 2.1 introduces the statistical assumptions on ε (the assumptions on X will be stipulated in Section 3). In Section 2.2, we introduce two naive estimators of the second moments of noise to motivate the ReMeDI approach.

2.1 The microstructure noise

Intuitively, microstructure noise emerges in each transaction (or quote), thus it is naturally a discrete process. The following assumptions on microstructure noise are satisfied by a broad spectrum of discrete processes.[4]

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

(1) $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}\left(\varepsilon_i^2\right) > 0 \quad \forall i \in \mathbb{Z}$;

(2) $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ *is a stationary and strongly mixing sequence with mixing coefficients* [5] $\{\alpha_k\}_{k=0}^{\infty}$, *and* $\alpha_k \downarrow 0$ *as* $k \to \infty$.

Remark 2.1. *There is a large literature seeking to characterize the economic mechanisms that govern the* dynamic properties *of microstructure noise, for example, the continuation of order flows modeled by Hasbrouck and Ho (1987), Choi et al. (1988), reversal order flows due to market maker's risk aversion by Grossman and Miller (1988) and Campbell et al. (1993) or inventory controls by Ho and Stoll (1981), Hendershott and Menkveld (2014), and the presence of inattentive (or infrequent) traders by Bogousslavsky (2016) and Hendershott et al. (2018). Econometric models of microstructure noise include i.i.d. process or an MA(q) process, see Hansen and Lunde (2006), Hautsch and Podolskij (2013) and Da and Xiu (2017), or ARMA(p, q) processes, see Barndorff-Nielsen et al. (2008), Hendershott et al. (2013). Note that the current settings of microstructure noise incorporate all the aforementioned models.*

Next, we impose some restrictions on the convergence rate of the mixing coefficients $\{\alpha_k\}_{k\in\mathbb{N}}$ that control the degree of serially dependence. In particular, the following assumption implies that the autocorrelation function of $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ is decaying at a polynomial rate.

[4] See Bradley (2005) for an excellent survey of the mixing sequences.

[5] 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 σ-algebra.

Assumption 2.2 (Polynomially mixing coefficients)**.** *There is some $C > 0, v > 0$ such that*

$$\alpha_k \leq \frac{C}{k^v} \quad \forall\, k \in \mathbb{N}^*. \tag{2}$$

Assumption 2.3. *Throughout the paper, we assume all moments of noise exist. This assumption can be relaxed depending on the targeted parameters and the choices of v in (2), see Lemma A.1. Consequently, an application of Lemma A.1 yields for some $C > 0$,*

$$\left|r_j\right| \leq \frac{C}{j^{v/2}}, \tag{3}$$

where $r_j = \mathbb{E}\left(\varepsilon_i \varepsilon_{i+j}\right)$, $\forall j \in \mathbb{N}$.

2.2 Two naive estimators of serial autocovariances

We begin with two estimators of serial autocovariances of stationary time series. Suppose the microstructure noise $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ satisfying Assumption 2.1 to 2.3 is *observable*. We would like to estimate the j-th order autocovariance $r_j = \mathbb{E}\left(\varepsilon_i \varepsilon_{i+j}\right)$ for some $j \in \mathbb{N}$.

2.2.1 The first naive estimator — the sample analogue

Given a sample of realized observations $\{\varepsilon_i\}_{1\leq i\leq n}$, the sample analogue provides a simple and intuitive estimator of r_j

$$\hat{r}_{n,j} := \frac{1}{n-j} \sum_{i=1}^{n-j} \varepsilon_i \varepsilon_{i+j}. \tag{4}$$

Under very mild conditions on the mixing coefficients $\{\alpha_k\}_{k\in\mathbb{N}^*}$, $\hat{r}_{n,j}$ has the following limit distribution for some $\Sigma_{\text{naive1}}(j) > 0$:[6]

$$\sqrt{n}\left(\hat{r}_{n,j} - r_j\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Sigma_{\text{naive1}}(j)\right). \tag{5}$$

2.2.2 The second naive estimator — a ReMeDI estimator

Now we introduce another consistent estimator. Let $\{k_n\}_{n\in\mathbb{N}^*}, \{k'_n\}_{n\in\mathbb{N}^*}$ be two sequences of integers that grow at slower rates than n: $k_n, k'_n \to \infty, k_n/n, k'_n/n \to 0$. The new estimator is given by

$$\text{ReMeDI}(\varepsilon; j)_n := \frac{1}{n-j-k_n-k'_n+1} \sum_{i=k'_n}^{n-j-k_n} \left(\varepsilon_i - \varepsilon_{i-k'_n}\right)\left(\varepsilon_{i+j} - \varepsilon_{i+j+k_n}\right). \tag{6}$$

Figure 1 depicts the estimators: $\text{ReMeDI}(\varepsilon; j)_n$ is the sample autocovariance of *disjoint increments over relatively large intervals*. $\text{ReMeDI}(\varepsilon; j)_n$ is also a consistent estimator of r_j. To get the intuition

[6] See Corollary 5.1 of Hall and Heyde (1980).

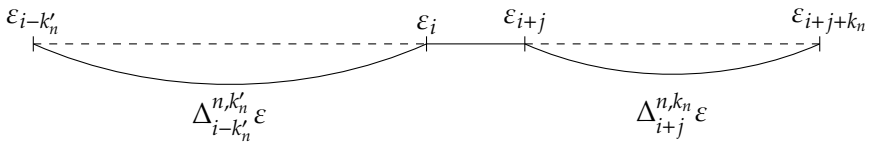

Figure 1: Illustration of the second naive estimator, $k_n, k_n' \to \infty$ as $n \to \infty$, but $k_n/n, k_n'/n \to 0$. Since $k_n, k_n' \to \infty$, the dependence of pairwise variables except $(\varepsilon_i, \varepsilon_{i+j})$ will shrink to zero since the "distances" between any two variables except $(\varepsilon_i, \varepsilon_{i+j})$ increase to infinity.

of its consistency, we first note

$$\left(\varepsilon_i - \varepsilon_{i-k_n'}\right)\left(\varepsilon_{i+j} - \varepsilon_{i+j+k_n}\right) = \varepsilon_i\varepsilon_{i+j} - \varepsilon_i\varepsilon_{i+j+k_n} - \varepsilon_{i-k_n'}\varepsilon_{i+j} + \varepsilon_{i-k_n'}\varepsilon_{i+j+k_n}.$$

The last three terms are the products of asymptotically independent and centered random variables, given $k_n, k_n' \to \infty$. Then (3) implies the expectations of the three terms are asymptotically negligible, thus their sample averages (over all i) will converge in probability to zero by law of large numbers. In the meanwhile, the sample average of $\varepsilon_i\varepsilon_{i+j}$, which is asymptotically equal to $\hat{r}_{n,j}$ (recall $k_n/n, k_n'/n \to 0$), converges in probability to r_j. Figure 1 also illustrates the intuition: the "distance" thus the independence of pairwise variables grows except for ε_i and ε_{i+j}. Under certain conditions[7], $\text{ReMeDI}(\varepsilon; j)_n$ has a limit distribution for some $\Sigma_{\text{naive2}}(j) > 0$:

$$\sqrt{n}\left(\text{ReMeDI}(\varepsilon; j)_n - r_j\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Sigma_{\text{naive2}}(j)\right). \tag{7}$$

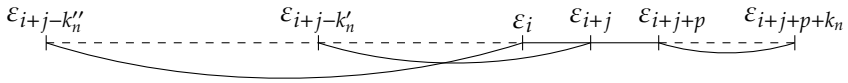

Figure 2: Illustration of the ReMeDI estimator of $\mathbb{E}\left(\varepsilon_0\varepsilon_j\varepsilon_{j+p}\right)$, $j, p \in \mathbb{N}$. $k_n, k_n', k_n'' \to \infty, k_n/n, k_n'/n, k_n''/n \to 0$ as $n \to \infty$.

$\text{ReMeDI}(\varepsilon; j)_n$ is essentially the realized autocovariances of disjoint increments (of microstructure noise), thus it is a special case of the ReMeDI class. Figure 2 illustrates a ReMeDI estimator of arbitrary third moment of noise (for any $j, p \in \mathbb{N}^*$) :

$$\text{ReMeDI}(\varepsilon; j, p)_n := \frac{1}{n - j - k_n - k_n' - k_n'' + 1} \sum_{i=k_n''}^{n-j-p-k_n} \left(\varepsilon_i - \varepsilon_{i-k_n''}\right)\left(\varepsilon_{i+j} - \varepsilon_{i+j-k_n'}\right)\left(\varepsilon_{i+j+p} - \varepsilon_{i+j+p+k_n}\right). \tag{8}$$

The same intuition for the consistency of $\text{ReMeDI}(\varepsilon; j)_n$ also applies here, thus one would expect that $\text{ReMeDI}(\varepsilon; j, p)_n$ provides a consistent estimator of $\mathbb{E}\left(\varepsilon_0\varepsilon_j\varepsilon_{j+p}\right)$.

[7] See Theorem D.1.

3 Separate Noise from Efficient Price by ReMeDI

The two naive estimators $\hat{r}_{n,j}$ and $\text{ReMeDI}(\varepsilon; j)_n$ are not applicable in practice since the microstructure noise is masked by the efficient price thus not directly observable. However, the second estimator $\text{ReMeDI}(\varepsilon; j)_n$ is still promising if ε is replaced by the observed price Y. In this section, we explain how the ReMeDI approach effectively separates microstructure noise from the efficient price. We consider two settings. First, without any specifications on the data frequencies, we demonstrate the intrinsic consistency of the ReMeDI estimators. The key of the separation is the martingale property of the efficient price and its independence of the microstructure noise. Next, under infill asymptotics, we show a more general consistency result and derive the limit distributions. The intuition behind this general separation principle is that microstructure noise dominates the variation of the efficient prices under infill asymptotics.

3.1 The intrinsic consistency of ReMeDI

Assume the efficient price $\{X_i\}_{i=0}^{\infty}$ is a discrete martingale. Then the observed price becomes

$$Y_i = X_i + \varepsilon_i, \quad i = 0, 1, 2, \dots \tag{9}$$

Now we introduce the *frequency-free* (FF) ReMeDI estimators of arbitrary second and third moments of noise. Given a sample of observed prices $\{Y_i\}_{i=0}^{n}$, for any $j, p, k_n \in \mathbb{N}^*$, let

$$\text{ReMeDI}(Y; j)_n^{\text{FF}} := -\frac{1}{n - 3k_n - j + 1} \sum_{i=2k_n}^{n-k_n-j} \Delta_{i+j}^{k_n} Y \Delta_{i-2k_n}^{2k_n} Y, \tag{10}$$

$$\text{ReMeDI}(Y; j, p)_n^{\text{FF}} := -\frac{1}{n - 4k_n - j - p + 1} \sum_{i=3k_n}^{n-k_n-j-p} \Delta_{i+j+p}^{k_n} Y \Delta_{i+j-2k_n}^{2k_n} Y \Delta_{i-3k_n}^{3k_n} Y, \tag{11}$$

where $\Delta_i^k Y = Y_{i+k} - Y_i$. One may note the analogy between $\text{ReMeDI}(Y; j)_n^{\text{FF}}$ and $\text{ReMeDI}(\varepsilon; j)_n$, $\text{ReMeDI}(Y; j, p)_n^{\text{FF}}$ and $\text{ReMeDI}(\varepsilon; j, p)_n$.

Assumption 3.1. *The efficient price $\{X_i\}_{i=0}^{\infty}$ and microstructure noise $\{\varepsilon_i\}_{i=0}^{\infty}$ satisfy the following*

(i) The efficient price process $\{X_i\}_{i=0}^{\infty}$ is a martingale with bounded fourth moments.

(ii) The noise process $\{\varepsilon_i\}_{i=0}^{\infty}$ satisfies Assumption 2.1 to 2.3 and is independent of X.

Theorem 3.1. *Let k_n, v satisfy*

$$k_n \to \infty,\ k_n/n \to 0 \text{ as } n \to \infty, \quad v > 2. \tag{12}$$

Under Assumption 3.1, we have

$$\text{ReMeDI}(Y; j)_n^{\text{FF}} \xrightarrow{\mathbb{P}} \mathbb{E}\big(\varepsilon_0 \varepsilon_j\big); \quad \text{ReMeDI}(Y; j, p)_n^{\text{FF}} \xrightarrow{\mathbb{P}} \mathbb{E}\big(\varepsilon_0 \varepsilon_j \varepsilon_{j+p}\big). \tag{13}$$

Proof. See Appendix B. □

3.2 Separation under infill asymptotics

The efficient price X is a Itô semimartingale defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ with the Grigelionis representation

$$X_t = X_0 + \int_0^t a_s \mathrm{d}s + \int_0^t \sigma_s \mathrm{d}W_s + \left(\delta \mathbf{1}_{\{|\delta|\leq 1\}}\right) \star (\mathfrak{p} - \mathfrak{q})_t + \left(\delta \mathbf{1}_{\{|\delta|>1\}}\right) \star \mathfrak{p}_t, \tag{14}$$

where W is a Brownian motion and $\mathfrak{p}$ is a Poisson random measure on $\mathbb{R}_+ \times \mathbb{R}$ with its compensator $\mathfrak{q}(\mathrm{d}t, \mathrm{d}x) = \mathrm{d}t \otimes \lambda(\mathrm{d}x)$ and λ is a σ-finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. One is referred to Aït-Sahalia and Jacod (2014) for the last two stochastic integrals with respect to random measures. Assumption 3.2 imposes some additional conditions that are satisfied for most continuous-time price processes in finance and econometrics.

Assumption 3.2. *The process a is optionally locally bounded, the process σ is adapted and càdlàg , δ is predictable and there is a localizing sequence (τ_n) of stopping times such that for each n and a deterministic nonnegative function J_n on $\mathbb{R}$ satisfying $\int J_n^2(x)\lambda(\mathrm{d}x) < \infty$ and such that $|\delta(\omega, t, x) \wedge 1| \leq J_n(x)$ for all (ω, t, x) with $t \leq \tau_n(\omega)$.*

Note that the efficient price X_t evolves according to *calendar time* t, whereas the microstructure noise ε_i is indexed by *transaction times* i (the i-th transaction). The two indexing schemes coincide after the realizations of observations. To see this, consider a fixed $t > 0$ and denote the transaction times by $i\Delta_n, i = 0, \ldots, N_t^n, X_i^n = X_{i\Delta_n}$, where Δ_n is the mesh of observation grid and $N_t^n = [t/\Delta_n]$ is the number of observations. We also denote $\varepsilon_i^n = \varepsilon_i$, $\forall 0 \leq i \leq N_t^n$. The transaction price is given by

$$Y_i^n = X_i^n + \varepsilon_i^n. \tag{15}$$

For any process V, we denote $\Delta_i^{n,k} V = V_{i+k}^n - V_i^n$. For any $j, p, q \in \mathbb{N}^*$, the high-frequency (HF) ReMeDI estimators of the moments of noise are given by

$$\mathrm{ReMeDI}(Y; j)_n^{\mathrm{HF}} := -\frac{1}{N_t^n - 3k_n - j + 1} \sum_{i=2k_n}^{N_t^n - k_n - j} \Delta_{i-2k_n}^{n,2k_n} Y \Delta_{i+j}^{n,k_n} Y, \tag{16}$$

$$\mathrm{ReMeDI}(Y; j, p)_n^{\mathrm{HF}} := -\frac{1}{N_t^n - 4k_n - j - p + 1} \sum_{i=3k_n}^{N_t^n - k_n - j - p} \Delta_{i-3k_n}^{n,3k_n} Y \Delta_{i+j-2k_n}^{n,2k_n} Y \Delta_{i+j+p}^{n,k_n} Y, \tag{17}$$

$$\mathrm{ReMeDI}(Y; j, p, q)_n^{\mathrm{HF}} := -\frac{\sum_{i=4k_n}^{N_t^n - k_n - j - p - q} \Delta_{i-4k_n}^{n,4k_n} Y \Delta_{i+j-3k_n}^{n,3k_n} Y \Delta_{i+j+p-2k_n}^{n,2k_n} Y \Delta_{i+j+p+q}^{n,k_n} Y}{N_t^n - 5k_n - j - p - q + 1}. \tag{18}$$

A few comments about the notations are in order. Under infill asymptotics, the mesh grid $\Delta_n \to 0$ as $n \to \infty$. In particular, the distribution of $X_{i\Delta_n}$ depends on n. Thus all variables in (15), as well as $\Delta_i^{n,k} V$ have an additional superscript n compared to their counterparts in Section 3.1.[8]

[8]The superscript n in ε_i^n merely indicates that $0 \leq i \leq N_t^n$, and it does not affect the statistical properties of ε.

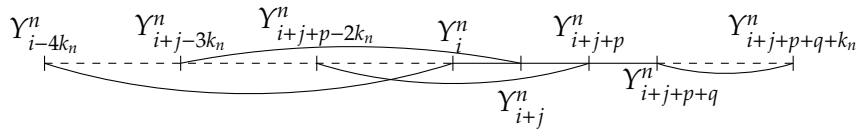

Figure 3: Illustration the ReMeDI estimator of $\mathbb{E}\left(\varepsilon_0\varepsilon_j\varepsilon_{j+p}\varepsilon_{j+p+q}\right)$, $j,p,q\in\mathbb{N}$.

The remaining part of this section presents the infill asymptotic properties of the high-frequency ReMeDI estimators. Theorem 3.2 and Theorem 3.3 provide the consistency and limit distribution. Compared to the consistency results in Section 3.1, the ReMeDI approach provides a consistent estimator of the fourth moments of noise under infill asymptotics and such consistency can not be achieved when the data frequency is fixed.[9] Theorem 3.4 introduces a robust estimator of the asymptotic variances thus makes the limit distribution *feasible* to construct confidence intervals.

Theorem 3.2. *Let the efficient price X be described in (14) and satisfy Assumption 3.2. The noise process ε satisfies Assumption 2.1 to 2.3 and k_n satisfies*

$$k_n\to\infty,\quad \Delta_n k_n\to 0. \tag{19}$$

We have the following consistency results for any $j,p,q\in\mathbb{N}$:

$$\text{ReMeDI}(Y;j)^{\text{HF}}_n\xrightarrow{\mathbb{P}}\mathbb{E}\left(\varepsilon_0\varepsilon_j\right); \tag{20}$$

$$\text{ReMeDI}(Y;j,p)^{\text{HF}}_n\xrightarrow{\mathbb{P}}\mathbb{E}\left(\varepsilon_0\varepsilon_j\varepsilon_{j+p}\right); \tag{21}$$

$$\text{ReMeDI}(Y;j,p,q)^{\text{HF}}_n\xrightarrow{\mathbb{P}}\mathbb{E}\left(\varepsilon_0\varepsilon_j\varepsilon_{j+p}\varepsilon_{j+p+q}\right). \tag{22}$$

Proof. See Appendix C. □

Note that we do not need the independence of X and ε to achieve the consistency.

Theorem 3.3 (Central Limit Theorem)**.** *Let the efficient price X be described in (14) and satisfy Assumption 3.2. The noise process ε satisfies Assumption 2.1 to 2.3, and it is also independent of X. Let k_n, v satisfy*

$$v>6,\quad k_n\asymp\Delta_n^{-\gamma},\gamma\in\left(\frac{1}{v},\delta\right),\text{where }\delta\in\left(\frac{2}{v+4},\frac{1}{5}\right). \tag{23}$$

Then we have the following convergence in distribution:

$$\sqrt{N^n_t}(\text{ReMeDI}(Y;j)^{\text{HF}}_n-r_j)\xrightarrow{\mathcal{L}}\mathcal{N}\left(0,\Sigma_j\right); \tag{24}$$

where

$$\Sigma_j:=\sum_{k=-\infty}^{\infty}\left(\mathbb{E}\left((\varepsilon_0\varepsilon_j-r_j)(\varepsilon_k\varepsilon_{k+j}-r_j)\right)+3r_k^2\right). \tag{25}$$

Proof. See Appendix D. □

[9] See Li and Linton (2018) for a more general treatment.

Remark 3.1. *The tuning parameter k_n is bounded both from above and blow. The lower bounded is to guarantee the ReMeDI estimators converge to the moments of noise at a rate faster than $\sqrt{\Delta_n}$, the upper bound is set to satisfy some Lindeberg condition.*

Theorem 3.4 (Feasible Central Limit Theorem)**.** *Let i_n satisfy*

$$i_n \asymp \Delta_n^{-1/5}. \tag{26}$$

Under the same conditions of Theorem 3.3, we have

$$\sqrt{N_t^n}\frac{\text{ReMeDI}(Y;j)_n^{\text{HF}} - r_j}{\sqrt{\widehat{\Sigma}(Y)_j^n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \tag{27}$$

where

$$\widehat{\Sigma}(Y)_j^n := \frac{1}{N_t^n - 3k_n - j - i_n + 1}\sum_{i=2k_n}^{N_t^n - k_n - i_n - j}\left(\left(\widetilde{Y}_i^n\right)^2 + 2\sum_{k=1}^{i_n}\widetilde{Y}_{i+k}^n\right); \tag{28}$$

$$\widetilde{Y}_i^n := -\Delta_{i-2k_n}^{n,2k_n} Y \Delta_{i+j}^{n,k_n} Y - \text{ReMeDI}(Y;j)_n^{\text{HF}}. \tag{29}$$

Proof. See Appendix E. □

4 New Measures of the Effective Bid-ask Spread

Measuring the effective spreads can be difficult for at least two reasons. First, market orders have complex dynamics, see Remark 2.1. To directly calculate the effective spread, one needs to reconstruct the entire orders and measure the deviation of average price from midquote. This is often, however, not possible for non-participants like researchers or financial regulators. Second, the data set of trading directions is not always available. Although algorithms, e.g., Lee and Ready (1991) and Ellis et al. (2000) are available to estimate the signs of transactions, it will inevitably bring more uncertainty.

Roll (1984) introduces a simple and elegant measure of effective bid-ask spread based on transaction price alone. However, the Roll measure is derived under several restrictive assumptions, e.g., constant spread, uncorrelated order flows, independence of efficient price and spread. Moreover, the large sample properties of the Roll measure is not available.[10]

We propose new measures of the effective spreads. The new measures are model free, and allow for serial autocorrelations in the spreads, correlation between the efficient price and spreads. The next section introduces the ReMeDI estimators of the new measures. The frequency-free (FF) and high-frequency (HF) asymptotic properties of the proposed estimators are also presented.

[10] Harris (1990) provides some simulation studies on the small sample properties of the Roll measure.

4.1 The Roll model: a revisit

To motivate the new measures, we revisit the classic Roll model of bid-ask spread. The efficient price $\{X_i\}_{i=1}^{\infty}$, captured by the midquote, follows a random walk. The bid-ask spread S is fixed. Thus the bid and ask prices are given by

$$a_i = X_i + \frac{S}{2}; \quad b_i = X_i - \frac{S}{2}.$$

The transaction prices are

$$Y_i = X_i + \frac{S}{2} q_i,$$

where $q_i = 1$ ($q_i = -1$) indicates buying (selling).

Remark 4.1. *Note that the* microstructure noise *is equivalent to the (signed) half spread that captures the deviation of transaction prices from the fundamental values, i.e.,* $\varepsilon_i = Sq_i/2$. *Assuming* $\mathbb{P}(q_i = 1) = \mathbb{P}(q_i = -1) = 1/2$, *the spread is essentially twice the standard deviation of noise:* $S = 2\sqrt{\mathbf{Var}(\varepsilon)}$.

4.2 An instantaneous bid-ask spread measure

We now introduce a new measure of effective bid-ask spread. Let

$$\varepsilon_i = S_i q_i / 2, \tag{30}$$

where S_i is the spread associated with the i-th trade and q_is are the the trading direction indicators with +1 for a buy, −1 for a sale. The bid and ask prices are $X_i + \frac{S_i}{2}$, $X_i - \frac{S_i}{2}$, respectively. Thus we can motivate ε as half of the signed bid-ask spread. Unlike Roll's settings, the effective spread S_i could to be random, which may vary, for example, with the size of trades. We allow for serial dependence in the trading directions, which are often found to be positively autocorrelated, see, e.g., Hasbrouck and Ho (1987), Huang and Stoll (1997), Sadka (2006) and Hendershott et al. (2013). Instead of modelling S_i, q_i separately, we directly impose some distributional assumptions (Assumptions 2.1) on $\varepsilon_i = S_i q_i/2$ to accommodate rich dynamics in spreads and order flows. Motivated by Remark 4.1, we introduce the *instantaneous bid-ask spread* (IBAS):

$$\text{IBAS} = 2\sqrt{\mathbf{Var}(\varepsilon)}.$$

4.3 An average bid-ask spread measure

The IBAS is based on the variance of ε only, thus it measures the *instantaneous price dispersion* — if a trader initiates a *single* trade, or several *sparse* trades, IBAS provides a reasonable measure of the associated spread. However, if an investor splits a large order into small pieces and sends the orders continuously over a long time period[11], he would concern more about the

[11] In the high-frequency setting, a *long* period could be just a few seconds.

average price dispersion that arises in the next hundreds or thousands of transactions, in which the autocorrelations in order flows can not be ignored. Such a measure, if it exists, serves as an accurate measure of the market quality as it gauges the average deviation of *all* transaction prices that occur in a period of time.

Proposition 4.1. *Let the signed half bid-ask spread satisfy Assumption 2.1 to 2.3, and the mixing coefficients* (α_k) *satisfy* $\sum_{k=1}^{\infty} \alpha_k^{\epsilon/(2+\epsilon)} < \infty$ *for some* $\epsilon > 0$. *Let* $\Gamma_\infty = \sum_{i=-\infty}^{\infty} r_i < \infty$, $\overline{\varepsilon}_n = \frac{\sum_{i=1}^{n} \varepsilon_i}{n}$. *Then*

$$\sqrt{n}\overline{\varepsilon}_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma_\infty).$$

One is referred to Corollary 5.1 in Hall and Heyde (1980) for a proof. Motivated by Remark 4.1 and Proposition 4.1, we define

Definition 4.1. *The* average bid-ask spread (ABAS) *is given by*

$$\text{ABAS} = 2\sqrt{\Gamma_\infty}. \tag{31}$$

5 Estimation of IBAS and ABAS through ReMeDI

5.1 Frequency-free estimators of IBAS and ABAS

First, we employ the frequency-free ReMeDI estimators introduced in Section 3.1 to construct the following estimators:

$$\text{IBAS}_n^{\text{FF}} := 2\sqrt{\text{ReMeDI}(Y;0)_n^{\text{FF}}}; \tag{32}$$

$$\text{ABAS}_n^{\text{FF}} := 2\sqrt{\left(\text{ReMeDI}(Y;0)_n^{\text{FF}} + 2\sum_{\ell=1}^{\ell_n^{\text{FF}}} \text{ReMeDI}(Y;\ell)_n^{\text{FF}}\right)}. \tag{33}$$

Theorem 5.1. *Under the assumptions of Theorem 3.1, we have*

$$\text{IBAS}_n^{\text{FF}} \xrightarrow{\mathbb{P}} \text{IBAS}. \tag{34}$$

If further we have $\ell_n^{\text{FF}}\sqrt{k_n/n} \to 0$, $\ell_n^{\text{FF}} k_n^{-v/2} \to 0$, *we have*

$$\text{ABAS}_n^{\text{FF}} \xrightarrow{\mathbb{P}} \text{ABAS}. \tag{35}$$

5.2 High frequency estimators of IBAS and ABAS

Given the transaction prices $\{Y_i^n\}_{0\le i\le N_t^n}$, the ReMeDI estimator of the IBAS under infill asymptotics is

$$\text{IBAS}_n^{\text{HF}} := 2\sqrt{\text{ReMeDI}(Y;0)_n^{\text{HF}}}. \tag{36}$$

Theorem 3.2 implies the following:

Proposition 5.1. *Let the efficient price X be described in (14) and satisfy Assumption 3.2. The signed half bid-ask spread ε satisfies Assumption 2.1 to 2.3 and k_n satisfies*

$$k_n \to \infty, \quad \Delta_n k_n \to 0.$$

We have the following consistency result :

$$\text{IBAS}_n^{\text{HF}} \xrightarrow{\mathbb{P}} \text{IBAS}. \tag{37}$$

Note that we do not assume the independence of X and ε to obtain the consistency result. By the delta method, Theorem 3.3 and Theorem 3.4, we have the limit distribution of $\text{IBAS}_n^{\text{HF}}$:

Theorem 5.2. *Let $X, \varepsilon, k_n, v, i_n$ satisfy the conditions in Theorem 3.3 and Theorem 3.4. Then*

$$\sqrt{N_t^n}\left(\text{IBAS}_n^{\text{HF}} - \text{IBAS}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Sigma_0/r_0\right), \tag{38}$$

and

$$\sqrt{N_t^n}\frac{\sqrt{\text{ReMeDI}(Y;0)_n^{\text{HF}}}\left(\text{IBAS}_n^{\text{HF}} - \text{IBAS}\right)}{\sqrt{\widehat{\Sigma}(Y)_0^n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{39}$$

To obtain an estimator of ABAS, we first obtain the following consistency result, which is a simple corollary of Theorem 3.2:

Corollary 5.1. *Given $v > 2$, let ι_n, k_n satisfy*

$$k_n \to \infty,\ \ell_n^{\text{HF}} \to \infty;\ \sqrt{k_n\Delta_n}\ell_n^{\text{HF}} \to 0,\ k_n^{-v/2}\ell_n^{\text{HF}} \to 0. \tag{40}$$

We have

$$\widehat{\Gamma}_n := \text{ReMeDI}(Y;0)_n^{\text{HF}} + 2\sum_{j=1}^{\ell_n^{\text{HF}}} \text{ReMeDI}(Y;j)_n^{\text{HF}} \xrightarrow{\mathbb{P}} \Gamma_\infty. \tag{41}$$

The high-frequency ReMeDI estimator of GABAS is

$$\text{ABAS}_n^{\text{HF}} := 2\sqrt{\widehat{\Gamma}_n}, \tag{42}$$

and (41) speaks to the consistency of $\text{ABAS}_n^{\text{HF}}$:

$$\text{ABAS}_n^{\text{HF}} \xrightarrow{\mathbb{P}} \text{ABAS}. \tag{43}$$

Deriving the asymptotic distribution of $\text{ABAS}_n^{\text{HF}}$ is also possible, but it is beyond the scope of this paper. We need to derive the *joint* distribution of the ReMeDI estimators that extends Theorem 3.3. This is studied in another companion paper by Li and Linton (2018).

5.3 Compare to the Roll measure

The Roll measure is related to the ReMeDI approach: both the Roll measure and the ReMeDI estimators of the second moments of noise are based on serial autocovariances of observed returns. In this subsection, we derive the asymptotic properties of the Roll measure and compare with the ReMeDI estimators. We demonstrate the inconsistency of the Roll measure when the order flows are serially correlated.

The fundamental value will be the midquotes and it follows a random walk. Let q_i be the trading direction of the i-th trade, $q_i = 1$ ($q_i = -1$) indicates buying (selling). Using our notations, the transaction prices are

$$Y_i = X_i + \frac{S}{2} q_i, \quad i = 0, 1, 2, \dots$$

The (frequency-free) Roll measure of S is given by

$$\text{Roll}_n^{\text{FF}} = 2\sqrt{-\frac{\sum_{i=1}^{n-1} \Delta_i^1 Y \Delta_{i-1}^1 Y}{n-1}}. \tag{44}$$

Now assume i) $(q_i)_{i\in\mathbb{N}}$ is i.i.d. with $\mathbb{E}(q_i) = 0$, ii) q is independent of X, iii) X has bounded fourth moments, we get the following by adapting the proof of Theorem 3.1:

$$\text{Roll}_n^{\text{FF}} \xrightarrow{\mathbb{P}} S. \tag{45}$$

If, however, the trading directions are autocorrelated with autocorrelation $\rho_q(i), i = 0, 1, 2, \dots$, then (45) becomes

$$\text{Roll}_n^{\text{FF}} \xrightarrow{\mathbb{P}} S\sqrt{1 - 2\rho_q(1) + \rho_q(2)}. \tag{46}$$

It is likely that the Roll measure becomes inconsistent[12]. However, we can apply the ReMeDI estimators to correct the bias terms in the Roll measure. The bias-adjusted Roll measure is thus given by

$$\text{AdjRoll}_n^{\text{FF}} = 2\sqrt{\left(\text{Roll}_n^{\text{FF}}/2\right)^2 + 2\text{ReMeDI}(Y;1)_n^{\text{FF}} - \text{ReMeDI}(Y;2)_n^{\text{FF}}}. \tag{47}$$

Consequently, we have

$$\text{AdjRoll}_n^{\text{FF}} \xrightarrow{\mathbb{P}} S.$$

Remark 5.1. *The inconsistency and bias correction of the* high-frequency *version of the Roll measure can be derived in a similar manner. In fact, the two versions of the Roll measures will return the same estimate* for a given sample. *Thus in the empirical analysis, we will stick to* Roll_n *and* AdjRoll_n

[12] Consider a simple model $\mathbb{P}(q_{i+1} = \pm 1 | q_i = \pm 1) = \pi$; $\mathbb{P}(q_{i+1} = \mp 1 | q_i = \pm 1) = 1 - \pi$, one can show that $\rho_q(1) = \rho$, $\rho_q(2) = \rho^2$ where $\rho = 2\pi - 1$. Then $1 - 2\rho_q(1) + \rho_q(2) = (1-\rho)^2$.

without explicitly mentioning the asymptotic framework.

We extend this notational convenience to the ReMeDI estimators as well in the sequel. Thus $\text{ReMeDI}(Y; j)_n$ *refers to either* $\text{ReMeDI}(Y; j)_n^{\text{FF}}$ *in* (10) *or* $\text{ReMeDI}(Y; j)_n^{\text{HF}}$ *in* (16).

6 Simulation Studies

This section presents extensive simulation studies to examine the performance of the ReMeDI estimators in finite samples. We introduce several leading benchmark models in financial econometrics and market microstructure to demonstrate the robustness of the ReMeDI approach from various perspectives. The model in Section 6.1 allows for jumps in both the efficient price and volatility level; we compare the ReMeDI estimators with the *local averaging* (LA) estimator recently proposed by Jacod et al. (2017). The structural model in Section 6.2 is based on Hasbrouck and Ho (1987), with an extension in the autocorrelation patterns of the bid-ask spreads. Section 6.3 studies a pricing error model introduced by Hendershott et al. (2013), in which the efficient price and noise (pricing errors) are correlated; the estimation is performed over samples of different data frequencies.

6.1 A statistical model

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

$$\begin{aligned}
&\mathrm{d}X_t = \kappa_1(\mu_1 - X_t)\mathrm{d}t + \sigma_t \mathrm{d}W_{1,t} + \xi_{1,t}\mathrm{d}N_t;\\
&\mathrm{d}\sigma_t^2 = \kappa_2(\mu_2 - \sigma_t^2)\mathrm{d}t + \gamma\sigma_t \mathrm{d}W_{2,t} + \xi_{2,t}\mathrm{d}N_t;\\
&\mathbf{Corr}(W_1, W_2) = \varrho;\\
&\xi_{1,t} \sim \mathcal{N}(0, \mu_2/10)\,;\; N_t \sim \text{Poi}(\lambda);\; \xi_{2,t} \sim \text{Exp}(\delta).
\end{aligned} \tag{48}$$

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

$$\kappa_1 = 0.5;\; \mu_1 = 1.6;\; \kappa_2 = 5/252;\; \mu_2 = 0.04/252;\; \gamma = 0.05/252;\; \varrho = -0.5;\; \lambda_1 = 5;\; \delta = \gamma.$$

We assume an AR(1) noise process, as studied in Aït-Sahalia et al. (2011):

$$\varepsilon_i = V_i + U_i, \tag{49}$$

where V is centered i.i.d. and U is an AR(1) process with first order coefficient ρ, $\left|\rho\right| < 1$. V and U are statistically independent. We set

$$\mathbf{Var}(V) = 2.9 \times 10^{-8};\; \mathbf{Var}(U) = 4.3 \times 10^{-8},\; \rho = 0.7.$$

Those estimates are borrowed from Aït-Sahalia et al. (2011). We tentatively set the AR(1) coefficient $\rho = 0.7$ to capture positive order flows find in literature. Later, we will vary the

choices of ρ to account for the complexity of noise dynamics, see a detailed discussion in Li et al. (2017).

6.1.1 Estimating autocovariances of noise

Figure 4 presents the estimation of the first 20 autocovariances of noise by ReMeDI (top panel) and LA (bottom panel). The solid lines are the mean estimates of 1,000 replications, the dashed lines represent the 95% simulated confidence intervals. We examine the performance of the estimators under different model specifications in which the price and/or volatility may exhibit jumps. We simulate 23,400 observations for each sample path, corresponding to the number of seconds of a business day (6.5 trading hours). The ReMeDI estimators perform well: the estimates are unbiased with compact confidence bands, which are slightly wider when the price and volatility processes exhibit jumps. Surprisingly, there is a significant deviation of the LA estimates to the true parameters. Moreover, the deviation and confidence bands become much larger when the price and/or volatility have jumps.

The deviation of the LA estimates is elicited by a *finite sample bias*, which is a fraction of the quadratic variation (QV) of the efficient price, see a discussion in Jacod et al. (2017). Thus to correct the bias, we need at least an estimate of the QV. But the estimation of QV in the presence of dependent noise is not trivial.[13] In a simulation context, we can obtain the QV thus can give the LA estimators the privilege to make the *exact* bias correction, which is, of course not feasible in practice. The bottom panel of Figure 5 displays the bias corrected estimation of LA. Even with the exact (yet infeasible) bias correction, the ReMeDI estimators still outperform the LA estimators with less bias and greater accuracy under all specifications.

6.1.2 Central limit theorem and finite sample distribution

To examine the limit distributions characterized in Theorem 3.3 and Theorem 3.4, we plot the quantiles of the normalized ReMeDI estimators $\sqrt{N_t^n}(\text{ReMeDI}(Y;j)_n - r_j)/\sqrt{\widehat{\Sigma}(Y)_j^n}$ against the quantiles of standard normal random variables. Figure 6 displays the plots with mesh grid fixed at $\Delta_n = 1$ sec. The plots well support the limit distributions. For practical concerns, it is desired to know how well the asymptotic variance estimator $\widehat{\Sigma}(Y)_j^n$ captures finite sample variations when the data frequencies are lower. Figure 7 replicates the plots in Figure 6 with $\Delta_n = 120$ sec. As expected the fits are not as perfect as in Figure 6, but $\widehat{\Sigma}(Y)_j^n$ only slightly underestimate the finite sample variance.

6.1.3 The choice of k_n

The tuning parameter k_n, which controls the length of the non-overlapping intervals to form the ReMeDI estimators, affects the performance of the estimators in finite samples. In this subsection, we propose several general rules to select k_n to improve the finite sample performance

[13] The realized volatility estimators of autocovariances of noise proposed by Li et al. (2017) also have a bias term induced by the QV. They propose a two-step approach to make the bias correction.

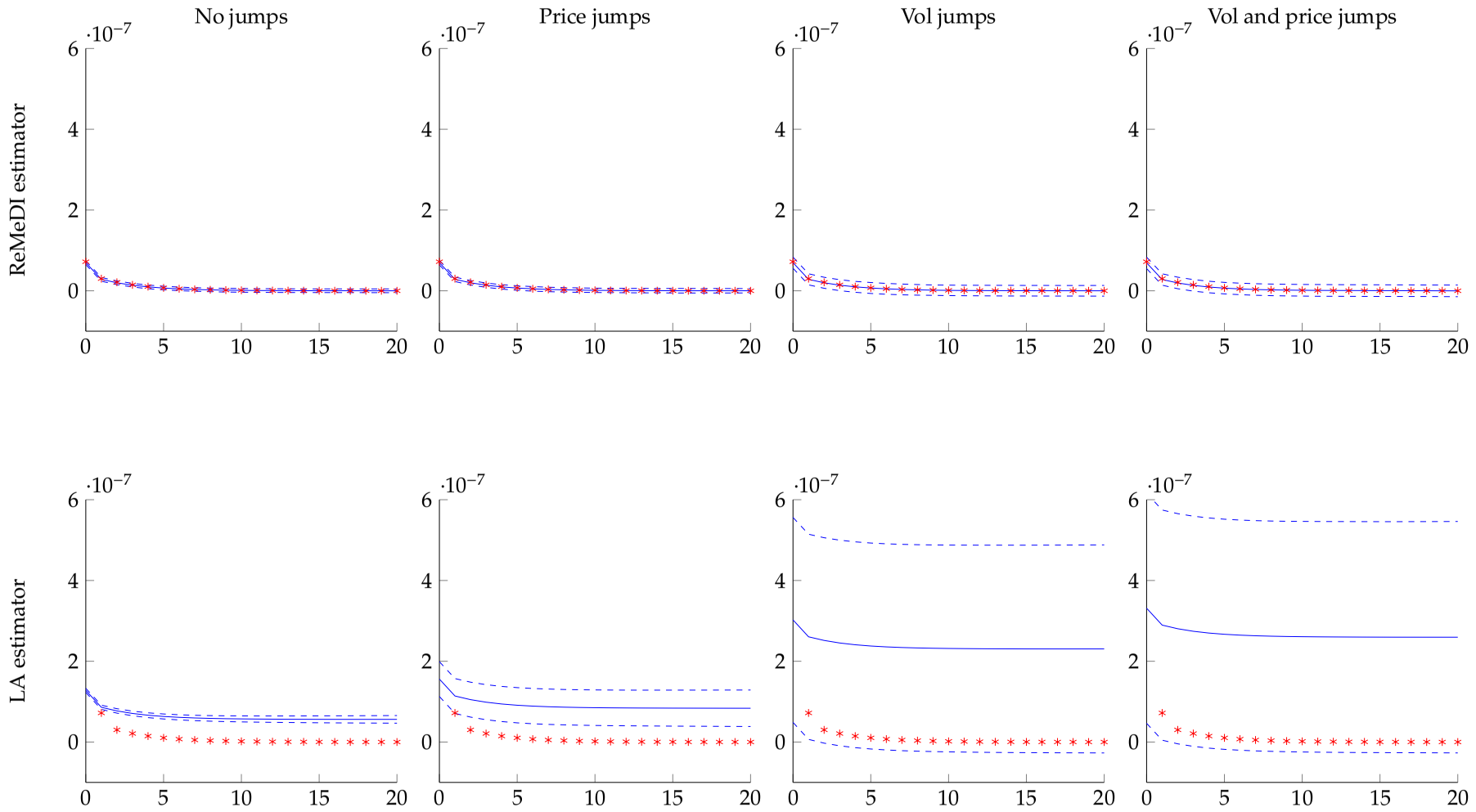

Figure 4: Estimation of autocovariances of microstructure noise using ReMeDI (top panel) and the local averaging (LA) method (bottom panel). The ReMeDI estimators are developed in (10) or (16). The model is specified in (48). The AR(1) coefficient $\rho = 0.7$, $\Delta_n = 1$ sec, $k_n = 10$, the tuning parameter of LA $k_n = 6$, number of replications is 1,000. The red stars are the true values of autocovariances of noise. The solid blue line is the mean estimates while the dashed lines are the simulated 95% confidence intervals. From left to right panels, the model specifications are continuous price and volatility, discontinuous price and continuous volatility, continuous price and discontinuous volatility, discontinuous price and volatility.

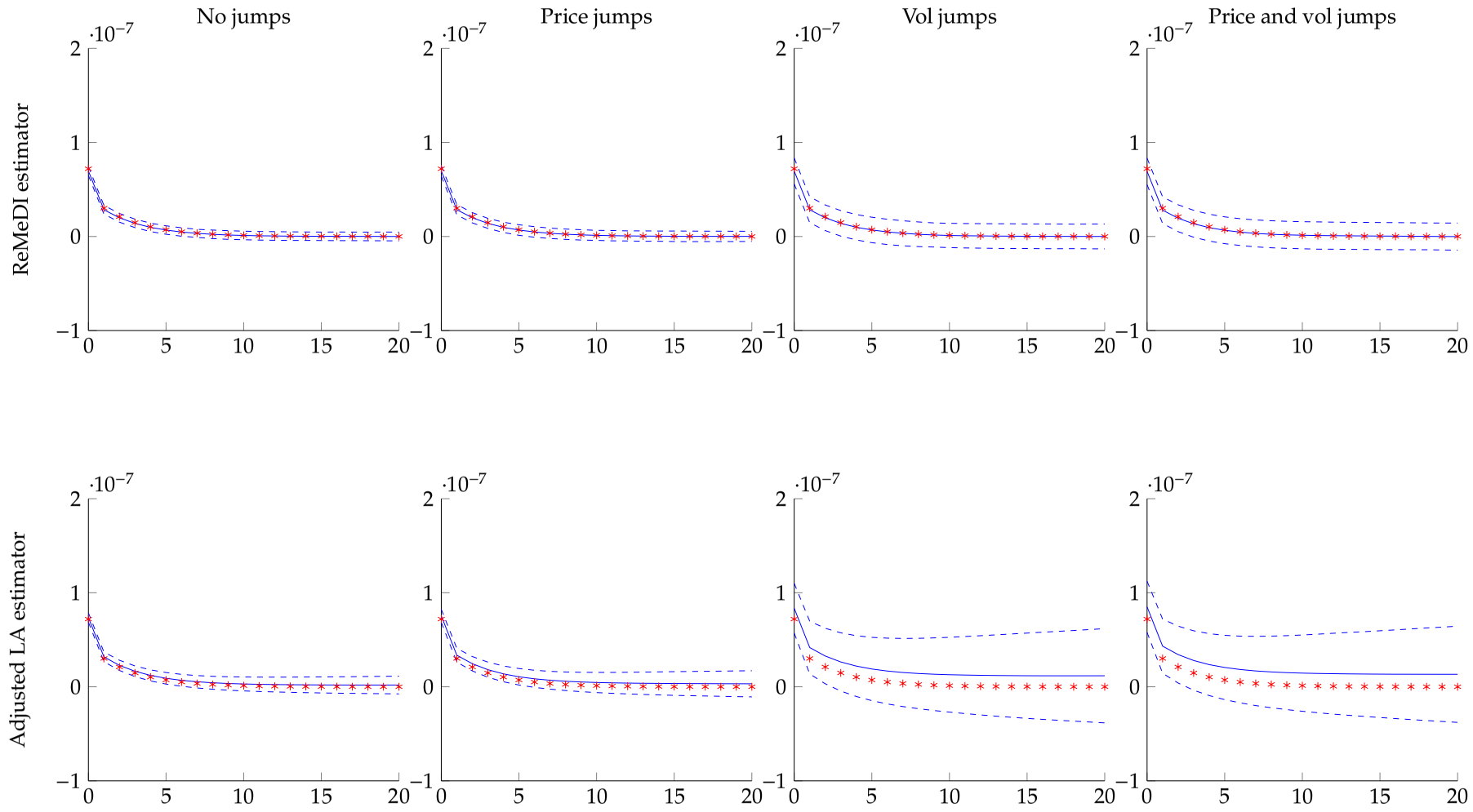

Figure 5: Estimation of autocovariances of microstructure noise using ReMeDI (top panel) and the exact bias corrected local averaging (LA) method (bottom panel). The ReMeDI estimators are developed in (10) or (16). The model is specified in (48). The AR(1) coefficient $\rho = 0.7$, $\Delta_n = 1$ sec, $k_n = 10$, the tuning parameter of LA $k_n = 6$, number of replications is 1,000. The red stars are the true values of autocovariances of noise. The solid blue line is the mean estimates while the dashed lines are the simulated 95% confidence intervals. From left to right panels, the model specifications are continuous price and volatility, discontinuous price and continuous volatility, continuous price and discontinuous volatility, discontinuous price and volatility.

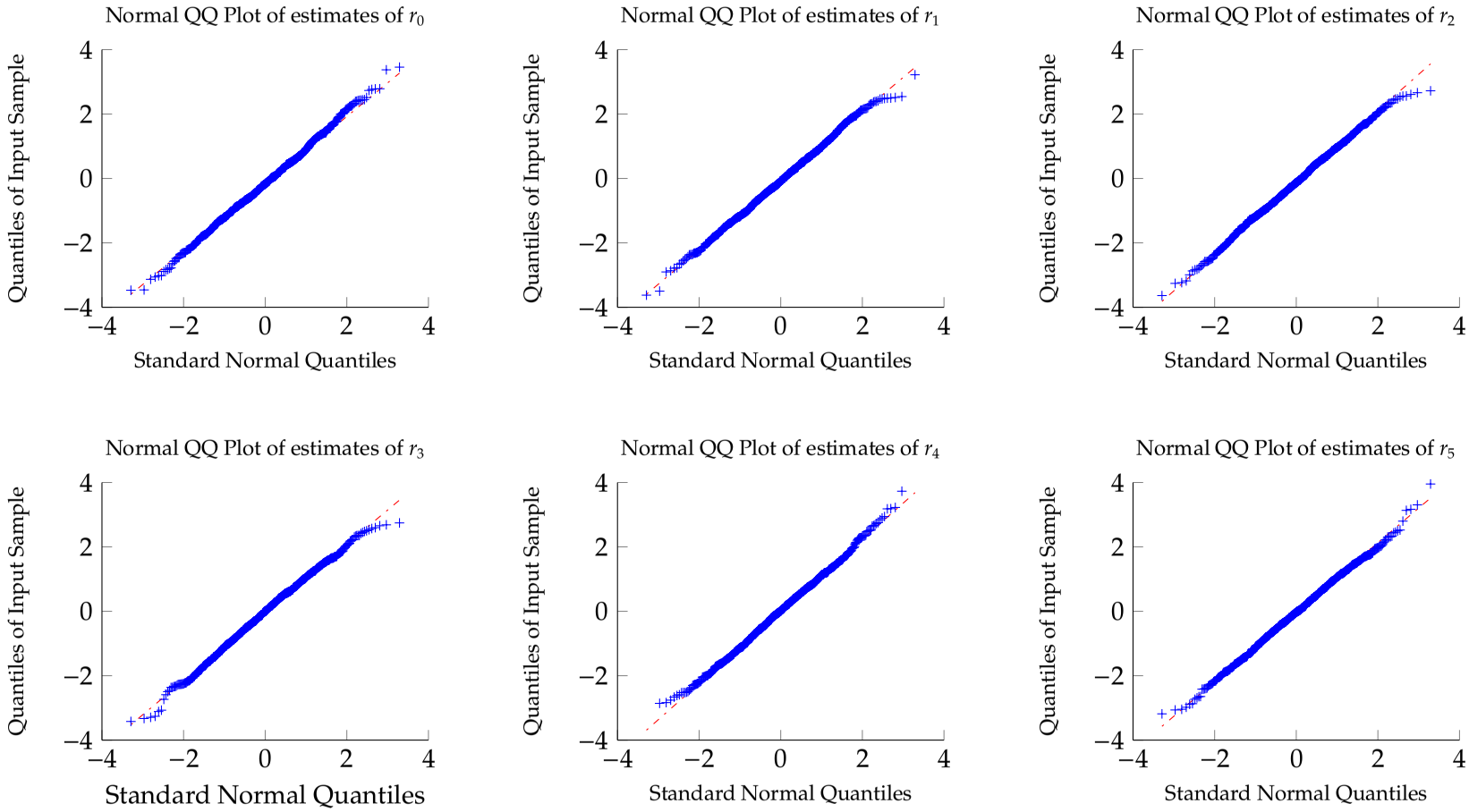

Figure 6: Standard normal QQ-plot of $\sqrt{N_t^n}\frac{\text{ReMeDI}((Y;j))_n - r_j}{\sqrt{\widehat{\Sigma}(Y)_j^n}}$. $\widehat{\Sigma}(Y)_j^n$ is calculated via (28). From left to right, top to bottom, $j = 0, 1, 2, 3, 4, 5$; respectively. The model is specified in (48). Number of simulation is 1,000. The tuning parameters $k_n = 10, i_n = 8$. $\Delta_n = 1$ sec.

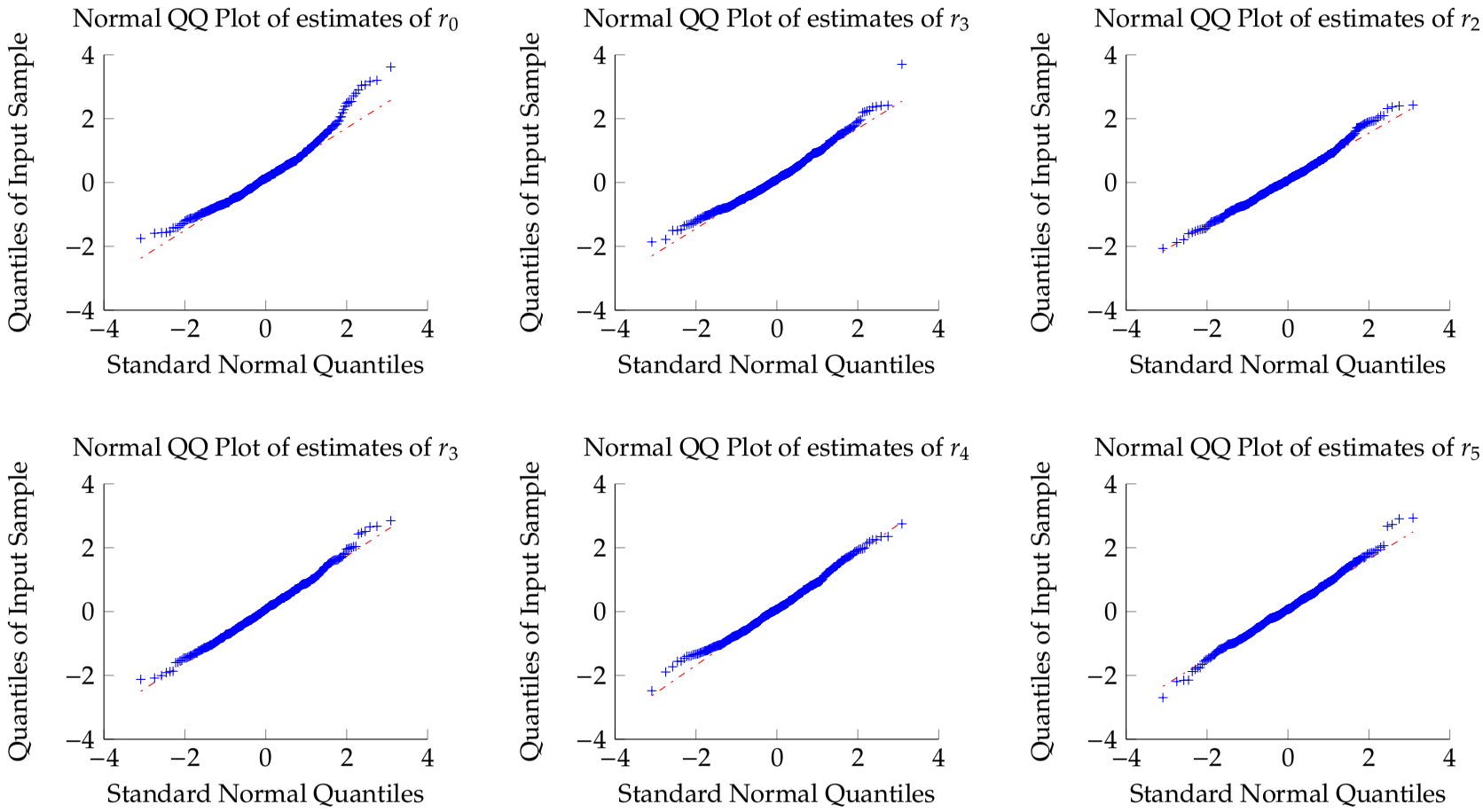

Figure 7: Standard normal QQ-plot of $\sqrt{N_t^n}\frac{\text{ReMeDI}((Y;j))_n - r_j}{\sqrt{\widehat{\Sigma}(Y)_j^n}}$. $\widehat{\Sigma}(Y)_j^n$ is calculated via (28). From left to right, top to bottom, $j = 0, 1, 2, 3, 4, 5$; respectively. The model is specified in (48). Number of simulation is 1,000. The tuning parameters $k_n = 10, i_n = 8$. $\Delta_n = 120$ sec.

based on numerical experiments.

We estimate the variance of noise with varying Δ_n and ρ (the AR(1) coefficient of noise) to replicate data samples at different frequencies with rich dynamic properties in the noise component. Under each specification, the bias, the standard deviation (std) and the root mean squared errors (RMSE) are reported to give to a full account of the finite sample performance of the ReMeDI estimators. All estimation results are collected in Table 1 to Table 3.

A close look at Table 1, Table 2 and Table 3 yields several observations. First, for each fixed k_n, the standard deviation and RMSE increase as the data frequency drops. This is intuitive since in a sparser sample, the efficient price entails more variation to the ReMeDI estimators, making it challenging to obtain accurate estimates. Second, the optimal k_n (in the RMSE sense) among the three candidates varies as the data frequency changes — a smaller k_n improves the performance when the data frequency is lower. Third, when the dependence in noise is weaker, a smaller k_n outperforms larger ones. Table 2 illustrates that for i.i.d. noise, the smallest k_n is always preferred regardless of the data frequencies.

	$k_n = 20$			$k_n = 10$			$k_n = 6$		
Δ_n	bias	std	RMSE	bias	std	RMSE	bias	std	RMSE
0.1s	-3.35e-11	1.08e-09	1.08e-09*	-1.25e-09	6.26e-10	1.40e-09	-5.63e-09	4.56e-10	5.65e-09
1s	-9.99e-10	2.91e-08	2.91e-08	-1.03e-09	1.15e-08	1.15e-08	-5.51e-09	5.48e-09	7.77e-09*
5s	6.98e-09	3.11e-07	3.11e-07	3.98e-09	1.16e-07	1.16e-07	-4.98e-09	5.44e-08	5.46e-08*
30s	5.12e-07	4.67e-06	4.70e-06	1.54e-07	1.70e-06	1.71e-06	5.63e-08	7.77e-07	7.79e-07*
60s	1.49e-06	1.30e-05	1.31e-05	3.37e-07	5.17e-06	5.18e-06	5.79e-08	2.13e-06	2.13e-06*

Table 1: Estimation of the variance of noise using the ReMeDI estimator. The model is specified in (48). The true value is 7.2×10^{-8}. Number of replications is 1,000. The AR(1) coefficient $\rho = 0.7$. The starred RMSE indicates the optimal k_n for each Δ_n.

	$k_n = 20$			$k_n = 10$			$k_n = 6$		
Δ_n	bias	std	RMSE	bias	std	RMSE	bias	std	RMSE
0.1s	1.14e-11	1.05e-09	1.05e-09	9.11e-12	4.83e-10	4.83e-10	3.93e-12	3.92e-10	3.92e-10*
1s	1.89e-09	2.88e-08	2.89e-08	-4.97e-10	1.07e-08	1.07e-08	-7.54e-11	5.15e-09	5.15e-09*
5s	-7.65e-09	3.30e-07	3.30e-07	4.13e-09	1.18e-07	1.18e-07	8.83e-10	5.55e-08	5.55e-08*
30s	4.23e-07	4.94e-06	4.96e-06	6.93e-08	1.64e-06	1.65e-06	3.58e-08	8.39e-07	8.40e-07*
60s	2.34e-06	1.29e-05	1.31e-05	3.39e-07	4.79e-06	4.80e-06	3.21e-07	2.33e-06	2.35e-06*

Table 2: Estimation of the variance of noise using the ReMeDI estimator. The model is specified in (48). The true value is 7.2×10^{-8}. Number of replications is 1,000. The AR(1) coefficient $\rho = 0$. The starred RMSE indicates the optimal k_n for each Δ_n.

	$k_n = 20$			$k_n = 10$			$k_n = 6$		
Δ_n	bias	std	RMSE	bias	std	RMSE	bias	std	RMSE
0.1s	-4.80e-11	1.09e-09	1.09e-09*	-1.28e-09	5.43e-10	1.39e-09	-5.56e-09	4.20e-10	5.58e-09
1s	2.32e-09	2.91e-08	2.92e-08	-8.19e-10	1.06e-08	1.06e-08	-5.32e-09	5.15e-09	7.40e-09*
5s	-3.52e-09	3.19e-07	3.19e-07	4.75e-09	1.08e-07	1.09e-07	-5.32e-09	5.38e-08	5.41e-08*
30s	1.35e-07	5.06e-06	5.06e-06	5.47e-08	1.64e-06	1.64e-06	4.85e-08	8.19e-07	8.20e-07*
60s	1.62e-06	1.29e-05	1.30e-05	4.63e-07	4.66e-06	4.68e-06	-2.21e-08	2.17e-06	2.17e-06*

Table 3: Estimation of the variance of noise using the ReMeDI estimator. The model is specified in (48). The true value is 7.2×10^{-8}. Number of replications is 1,000. The AR(1) coefficient $\rho = -0.7$. The starred RMSE indicates the optimal k_n for each Δ_n.

6.2 A generalized Hasbrouck and Ho (1987) model

Hasbrouck and Ho (1987) introduce a model of lagged price adjustment with positively auto-

correlated bid-ask spread. We follow their approach and model the efficient price by

$$X_t = \sigma W_t.$$

The midpoint of bid and ask is given by

$$p^n_{i+1} = p^n_i + \delta(X^n_i - p^n_i),$$

where $\delta \in [0,1]$ is a parameter to reflect the price adjustment: quotes may lag efficient price due to transaction costs. Compared to the original model in Hasbrouck and Ho (1987), a distinct feature of the model presented here is that we explicitly model the data/sampling frequencies, this is reflected in the superscript n in our notations (recall $X^n_i = X_{i\Delta_n} = \sigma W_{i\Delta_n}$). Frequency matters specifically in the study of the dynamic properties of microstructure effects, see, e.g., Hasbrouck and Sofianos (1993). Therefore any practice that alter the frequency of the examined data, for example, subsampling as Hasbrouck and Ho (1987) did will affect the statistical inference of the components of bid-ask spread and the distribution of returns.

Let ε be the half bid-ask spread so that actual transaction price is given by

$$Y^n_i = p^n_i + \varepsilon^n_i.$$

In Hasbrouck and Ho (1987), ε is an AR(1) process so that it can capture the serial dependence in buy and sell orders. As a consequence, the observed returns will be an ARMA(1,1) process. In the sequel, we assume ε is an ARMA(p,q) process (without specifying p,q) with Gaussian innovations[14] so that we have great flexibility in modeling the dynamics of order flows and distributions of observed returns:

$$\varepsilon_i = e_i + \sum_{j=1}^{p} \rho_j \varepsilon_{i-j} + \sum_{j=1}^{q} \gamma_j e_{i-j}, \quad e_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_e^2\right). \tag{50}$$

However, our estimators of the parameters of ε are essentially nonparametric, we do not invoke any parametric information of the bid-ask spread in this model. In the sequel, we borrow the empirical estimates from Bandi et al. (2017) and set

$$\delta = 0.01, \quad \sigma_e^2 = 1.9 \times 10^{-4}, \quad \sigma = 9.3 \times 10^{-3}.$$

6.2.1 Patterns of bid-ask spread and the distribution of returns

The aim of this subsection is to demonstrate that it could be misleading to make inference of noise from observed returns. Two contrasting noise processes with random walk efficient prices could generate return processes that are difficult to disentangle.

If the bid-ask spread is absent and price adjustment is instant ($\delta = 1$), the observed returns

[14] If the distribution of innovations is absolutely continuous with respect to Lebesgue measure, the ARMA process is a strongly mixing sequence, see Mokkadem (1988).

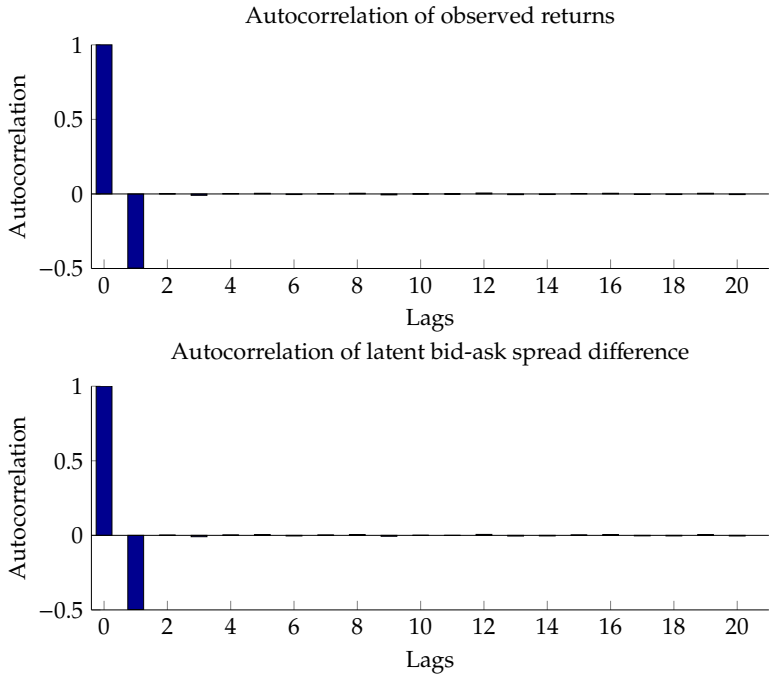

Figure 8: Autocorrelation function (ACF) of observed returns and latent noise (bid-ask spread) difference. The noise follows an i.i.d. process with variance σ_ε^2. The top panel plots the ACF of noisy returns and the bottom panel plots the ACF of noise differences. $\Delta_n = 1$ sec.

will be uncorrelated; otherwise the patterns of bid-ask spread will affect the distribution of observed returns when the data frequency is relatively high[15]. Very often stock returns have a significantly negative first order autocorrelation, the celebrated Roll's model provides a possible explanation : if the martingale efficient price and i.i.d. bid-ask spread constitute the observed price, the observed returns will become an MA(1) process. Figure 8 plots the autocorrelation function (ACF) of the observed returns when the bid-ask spread follows an i.i.d. process.

However, we should keep in mind that an MA(1) pattern of returns is a *consequence* but not the *cause* of an i.i.d. bid-ask spread. Figure 9 depicts the ACF of a return process that is virtually identical to the ACF in Figure 8, but the latent bid-ask spread follows an ARMA(1,1) process. Therefore one should be wary to conduct inference about the bid-ask spread based on the distribution of returns. In particular, assuming an i.i.d. bid-ask spread by observing MA(1) returns could oversimplify the dynamics of the underlying bid-ask spread. Nevertheless, the ReMeDI estimators provide a robust solution to deal with such complexities.

6.2.2 Estimating bid-ask spread

Now we employ the ReMeDI estimators to recover the dynamic properties of bid-ask spread. We plug in several sets of parameters to the ARMA(p, q) models specified in (50) to accommodate rich structures. Figure 10 provides the estimation results of the variance and autocovari-

[15]When the data frequency is high, the bid-ask spread dominates the efficient price process thus the second moments of observed returns are largely attributed to the second moments of the (first differences of) bid-ask spread, see Figure 8 and 9. The price adjustments will not affect the second moments of observed returns.

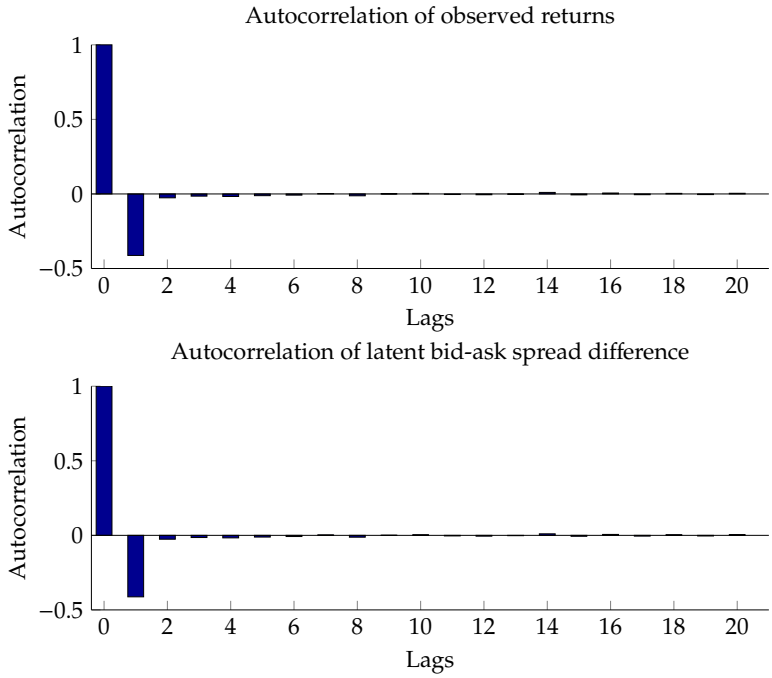

Figure 9: Autocorrelation function (ACF) of observed returns and latent noise (bid-ask spread) difference. The noise follows an ARMA(1,1) process: $\varepsilon_i = \rho_1 \varepsilon_{i-1} + e_i + \gamma_1 e_{i-1}$ with $\rho_1 = 0.7, \gamma_1 = -0.4$. The top panel plots the ACF of noisy returns and the bottom panel plots the ACF of noise differences. $\Delta_n = 1$ sec.

ances of bid-ask spread generated by 6 models. Clearly the results demonstrate the robustness of the ReMeDI estimators to different model specifications. Of particular interest are the i.i.d. (left top panel) and ARMA(1,1) (left bottom panel) models — if the efficient price is masked by bid-ask spread generated by the two models, the observed returns are close to MA(1) processes that are hard to distinguish, as illustrated in Figure 8 and 9 — the ReMeDI estimators, however, are able to disentangle the underlying bid-ask patterns.

6.3 The Hendershott et al. (2013) model of pricing error

To estimate the implementation shortfall, Hendershott et al. (2013) introduce a model that decompose the price process into permanent efficient price and transitory pricing error. They argue that the transitory pricing errors would be very persistent when investors split large orders into smaller ones and feed them into the market gradually. They model such persistence by an AR(2) process. Using our notations, their model becomes

$$\begin{cases} Y_i^n = X_i^n + \varepsilon_i^n, \\ X_{i+1}^n = X_i^n + w_i^n, \\ \varepsilon_{i+1} = \psi_1 \varepsilon_i + \psi_2 \varepsilon_{i-1} + e_i. \end{cases} \quad \begin{cases} w_i^n \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \Delta_n \sigma_w^2\right), \\ e_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_e^2\right), \\ \mathbf{Cov}\left(w_i^n, e_i\right) = \sqrt{\Delta_n} \rho \sigma_e \sigma_w. \end{cases} \tag{51}$$

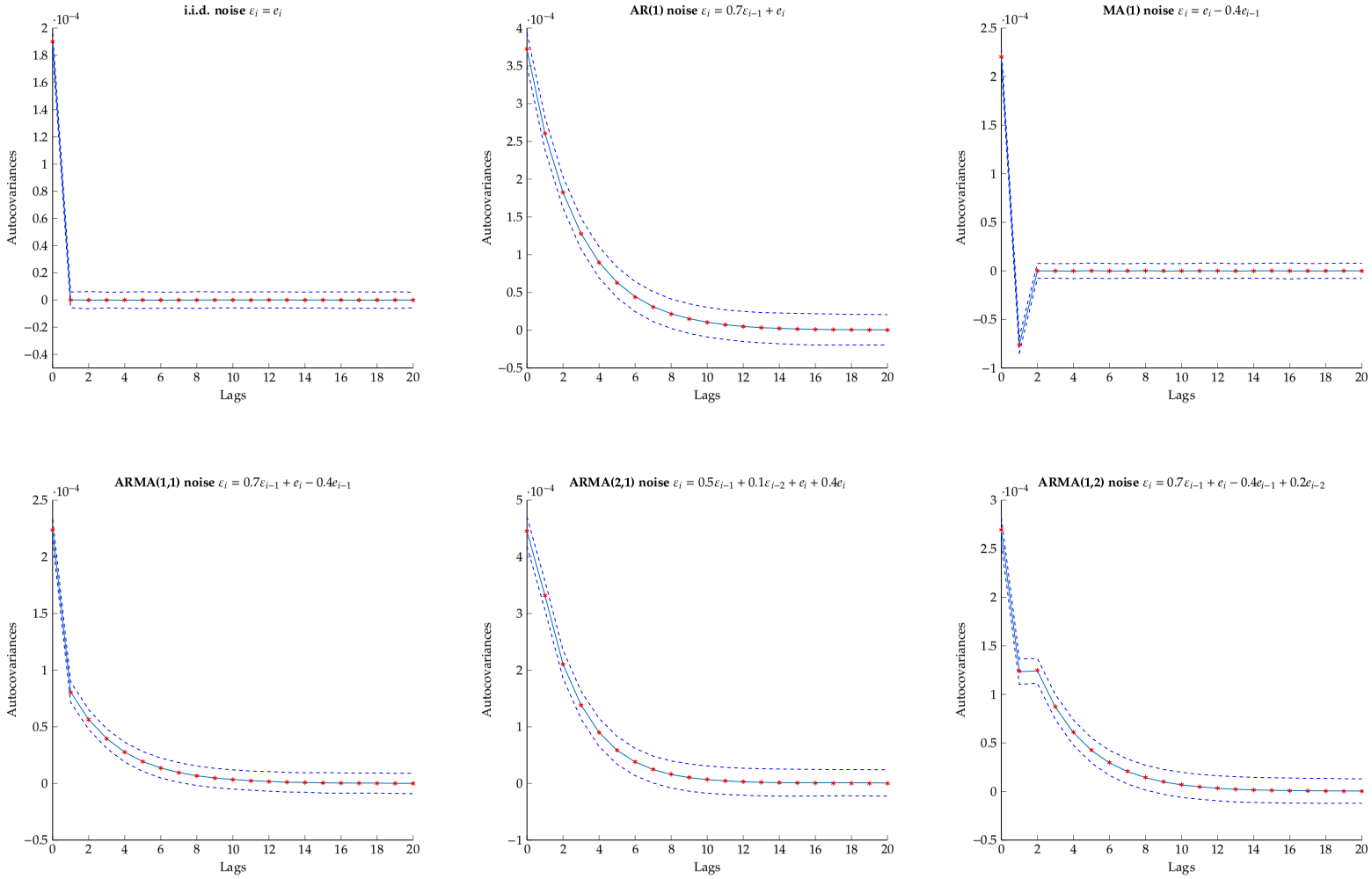

Figure 10: Estimation of autocovariances of bid-ask spread using ReMeDI estimators. The model is specified in (50). $k_n = 20, \Delta_n = 1$ sec, number of replications is 1,000. The red stars are the true values of autocovariances of noise. The solid blue line is the mean estimates while the dashed lines are the simulated 95% confidence intervals.

Note that they allow the efficient price to be correlated with pricing errors to incorporate informational effect. We use the estimates in Hendershott et al. (2013) and set

$$\sigma_w = 9.1 \times 10^{-3},\ \sigma_e = 4.4 \times 10^{-3},\ \rho = 0.38,\ \psi_1 = 0.65,\ \psi_2 = -0.05.$$

The inter-observation lag Δ_n ranges from 1 second to 10 minutes.

We can recover the parameters of pricing errors using the ReMeDI estimators. The results are presented in Figure 11. The blue solid lines are the average estimates of the first 20 autocovariances based on 1,000 samples. We observe that the confidence bands become wider as the data frequency shrinks — this is intuitive since the variation of the permanent efficient price dominates the pricing errors when the frequency is lower. Nevertheless, the ReMeDI estimators retain great accuracy with almost negligible bias.

7 Finite Sample Analysis

This section explains the finite sample robustness of the ReMeDI design presented in previous section. We show the ReMeDI estimators have a very small bias term and its magnitude is not affected by other variables or parameters. We illustrate that the confidence intervals constructed from the limit distribution is robust to data frequencies. We explain why the ReMeDI estimators are robust to jumps. We propose several heuristic rules to select the tuning parameter to improve the finite sample performance.

7.1 Finite sample bias

Let the efficient price process X be a martingale, and assume the increment of X and the increment of noise on non-overlapping intervals are uncorrelated. The bias of the ReMeDI estimator of the second moments of noise is then given by

$$\text{Bias} = \mathbb{E}(\text{ReMeDI}(Y; j)_n) - r_j = r_{j+2k_n} - 2r_{j+k_n}. \tag{52}$$

Note that the bias term only depends on the parameters of noise $r_{j+2k_n} - 2r_{j+k_n}$, which is typically much smaller than the targeted parameter r_j for large k_n. The local averaging estimator, in contrast, have bias terms that stem from the integrated volatility of the efficient price. Li et al. (2017) show that the bias terms could be larger than the noise parameters of interest in practical implementations, they also show that without correcting the bias terms, one would obtain very misleading estimates of the noise parameters and the integrated volatility of the efficient price process.

Li et al. (2017) propose a two-step approach to solve the intrinsic bias problem: In the first step, they obtain an estimate of the integrated volatility, and this estimate will be employed to correct the bias in the second step. Our ReMeDI estimator takes a different approach: by taking the realize moments on non-overlapping intervals, we effectively remove the bias due to the efficient price. The validity of this approach hinges on the *martingale property* of efficient price,

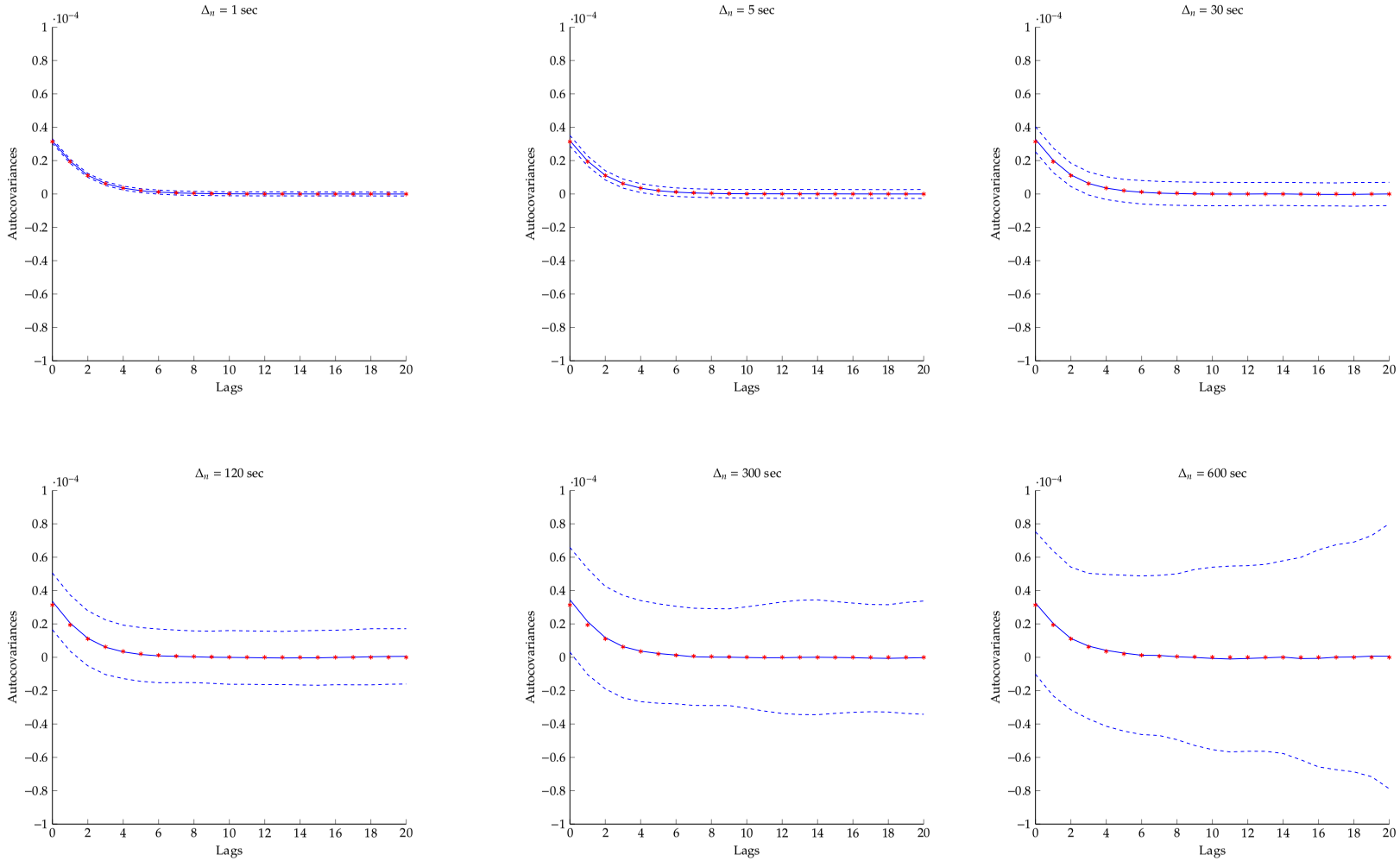

Figure 11: Estimation of autocovariances of pricing errors using ReMeDI estimators. The model is specified in (51). The number of replications is 1,000. The red stars are the true values of the autocovariances. The solid blue line is the mean estimates while the dashed lines are the simulated 95% confidence intervals. The parameters are $\sigma_w = 9.1 \times 10^{-3}$, $\sigma_e = 4.4 \times 10^{-3}$, $\rho = 0.38$, $\psi_1 = 0.65$, $\psi_2 = -0.05$. The data frequency Δ_n are, from left to right, top to bottom, 1 sec, 5 sec, 30 sec, 120 sec, 300 sec, 600 sec; respectively. The tuning parameters k_n are set, for each specification, $16, 12, 10, 8, 6, 4$; respectively.

which is implied by the Efficient Market Hypothesis, a cornerstone in modern finance theory. This highlights the economic intuition behind the ReMeDI estimators: The non-overlapping efficient returns "cancel out" and what remains is due to market frictions; by varying the "distances" and "sizes" of the non-overlapping windows, we can freely estimate the targeted parameters of microstructure noise. Therefore the ReMeDI approach greatly improves finite sample performance in ways that would otherwise invoke the estimation of the efficient price parameters to correct bias.

7.2 Finite sample distribution

Theorem 3.3 and Theorem 3.4 characterize the limit distribution of the ReMeDI estimators under infill asymptotics. Two concerns arise to construct confidence intervals using the limit distribution. First, we need a consistent estimator of the asymptotic variance. Second, among all available estimators if there is any, we would like to select the one that well measures the finite sample variance.

The proposed asymptotic variance estimator $\widehat{\Sigma}(Y)^n_j$ in (28) is not the only consistent estimator. The following is an alternative, and it is constructed using the ReMeDI estimators of the second and fourth moments of noise:

$$\begin{aligned}\widehat{\Sigma}(Y)'^n_j :=& \text{ReMeDI}(Y;0,j,0)^{\text{HF}}_n \\ &+ 2\sum_{i=1}^{i_n}\left(\text{ReMeDI}(Y;\min\{i,j\},|i-j|,\min\{i,j\})^{\text{HF}}_n + 3\left(\text{ReMeDI}(Y;i)^{\text{HF}}_n\right)^2\right) \\ &- (2i_n+1)\left(\text{ReMeDI}(Y;j)^{\text{HF}}_n\right)^2.\end{aligned}$$

Indeed, $\widehat{\Sigma}(Y)'^n_j$ is an excellent estimator of Σ_j. However, Σ_j itself is a poor measure of the finite sample variance of the ReMeDI estimators with relatively low data-frequency. It fails to take into account the variance of the efficient price process, which though is asymptotically negligible. As a consequence, $\widehat{\Sigma}(Y)'^n_j$ tends to underestimate the finite sample variance thus overstate the accuracy of the the ReMeDI estimators. While $\widehat{\Sigma}(Y)^n_j$ takes the form of sample analogue of the asymptotic variance Σ_j, it is also close to the finite sample variance. As showed in the simulation section, $\widehat{\Sigma}(Y)^n_j$ provides reasonable estimates of the finite sample variance in samples with data frequency up to several minutes.

7.3 Robustness to jumps

The ReMeDI estimators are also robust to abrupt extreme events, referred to as *jumps* in financial econometrics. The intuition of the robustness is rooted in *bipower* type estimators of the integrated volatility, see, among others, Barndorff-Nielsen and Shephard (2004), Barndorff-Nielsen and Shephard (2006) — taking the realized moments on disjoint intervals mitigate the impacts of jumps. It is worth mentioning that the presence of jumps will not affect the *infill asymptotic* properties of our ReMeDI estimators since the noise component has larger

asymptotic orders than jumps. However, in any finite samples where the magnitude of noise is in line with the size of jumps, the ReMeDI approach retains its accuracy, see Section 6.1.

7.4 Finite sample performance and the choice of k_n

It is temping to conclude from (52) that a large k_n is always preferred to reduce the finite sample bias. However, optimizing the finite sample performance by opting for large k_n is not always desired, as it fails to account for the *variance* of the ReMeDI estimators, which is partially contributed by the efficient price. Intuitively, a large k_n, thus large non-overlapping intervals to form the ReMeDI estimators triggers larger volatility attributed to the efficient prices, as the volatility is proportional to the length of the time span.

Deriving an analytical form of the optimal choice of k_n (in the mean squared error sense) is beyond the scope of this paper. Nevertheless, there are some heuristic rules on the selection of k_n in applications. If the dependence of noise is weak, a smaller k_n is preferred. In practice one can get some preliminary estimates of the autocorrelation functions of noise using the ReMeDI estimators with some arbitrarily chosen k_n, then repeat the estimation with a smaller k_n if the autocorrelation functions is decaying rapidly. Data frequency also affects the choice of k_n. In a sample with lower data frequency, a smaller k_n is preferred to reduce the volatility caused by the efficient price as the efficient price is dominating the noise component in a sparse sample. Some practical rules on the choice of k_n are also provided in the simulation studies, see Section 6.1.

8 Empirical Studies

8.1 Data description

We obtain the transaction prices from the TAQ data set for the Intel Corporation (INTC, Nasdaq) and the Coca-Cola Company (KO, NYSE) over the month February, 2016 (20 trading days). We remove observations prior to 9:30 AM and after 4:00 PM. Table 4 presents the summary statistics of the two stocks.

Company	Mean Price	Std Price	Mean Volume	Std Volume	Transactions/sec
INTC	29.14	0.63	194.51	2439	5.17
KO	43.04	0.54	177.11	1689	3.31

Table 4: Descriptive statistics for transaction price and trading volume of INTC and KO over February, 2016. The means and standard deviations of price and trading volume are calculated using all transactions over the trading month.

8.2 Autocorrelation patterns of noise

We first recover the autocorrelation patterns of microstructure noise using the ReMeDI estimators. We restrict our attention to the transactions after 9:35 AM, and leave the study of market opening to later sections.

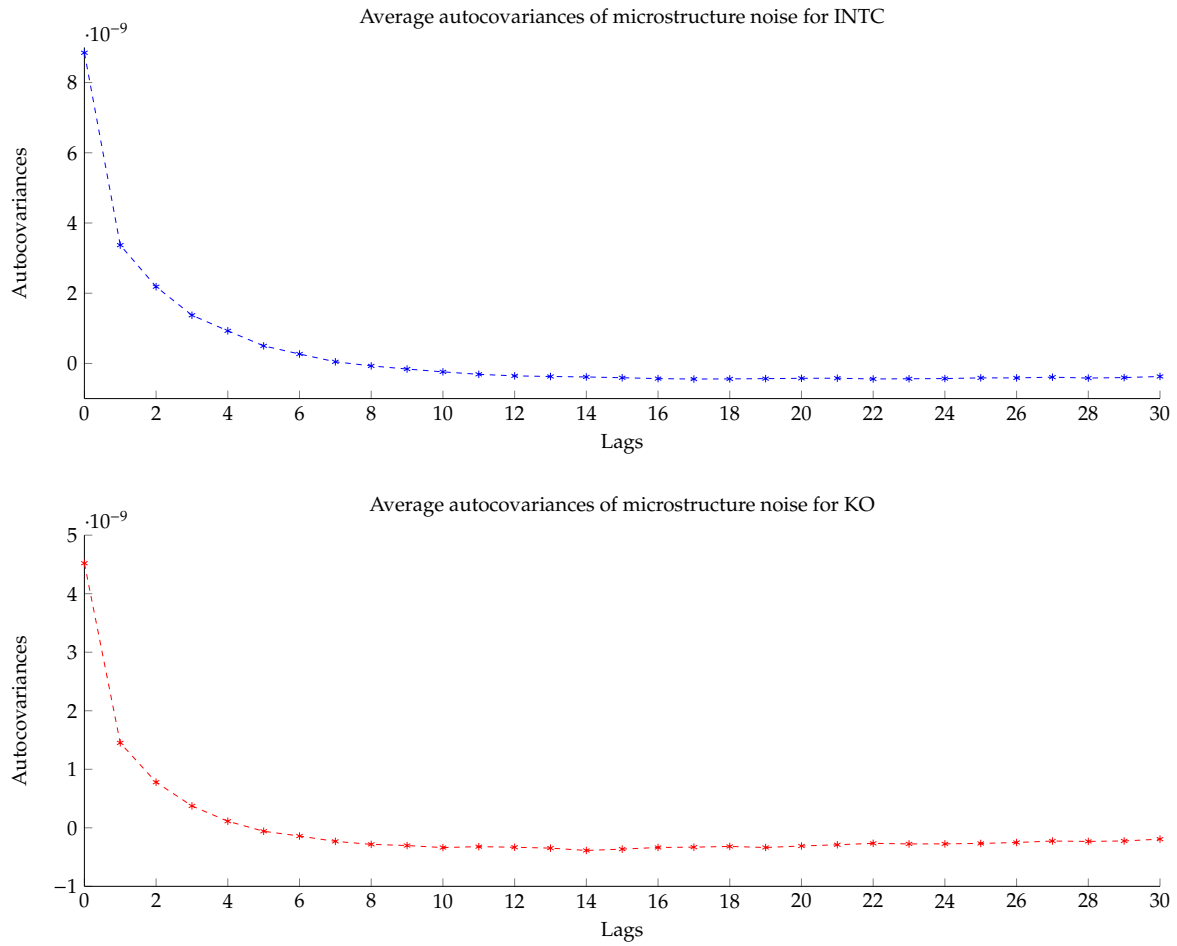

Figure 12: Average autocovariances of microstructure noise for INTC (top panel) and KO (bottom panel). In each of the 20 trading days of February, 2016, the ReMeDI estimates (with $k_n = 10$) of the autocovariances (up to 30 lags) are obtained using transaction prices between 9:35 AM and 4:00 PM. The mean of the autocovariances over the 20 trading days is plotted.

Figure 12 plots the mean estimates of the variance and autocovariances (up to 30 lags) over the 20 trading days.[16] It is clear that the microstructure noise for both stocks are positively autocorrelated. An intuitive interpretation is that large orders are often split into smaller ones and sent to the market gradually. Limit order can also generate positive autocorrelated transaction flows: when dealers change their quotes, only the stale orders on one side of the order book are executed. In either scenarios, these orders are recorded as separate transactions. Consequently, the effective spreads are positively autocorrelated, see Hasbrouck and Ho (1987), Choi et al. (1988), Hendershott et al. (2013) and Li et al. (2017), among others.

8.3 Measuring the bid-ask spread

We apply the estimators developed in Section 5 to measure the instantaneous bid-ask spread (IBAS) and average bid-ask spread (ABAS), and compare with the classic Roll measure.

The analysis in Section 5.3 reveals that the Roll measure and the ReMeDI measure of IBAS IBAS_n will coincide with each other were the order flows uncorrelated. Figure 13 presents

[16] In the appendix, we present the daily estimates, see related findings in Figure 18 and Figure 19.

the daily estimates of the two measures. Two observations are immediate. First, the two measures for INTC are slightly larger than that of KO. This is consistent with earlier empirical studies that spreads are higher in Nasdaq than NYSE, see, e.g., Stoll (2000, 2003). Second, there is a persistent and significant discrepancy between the two measures for both stocks. Such discrepancy is attributed to the nontrivial autocorrelation patterns of order flows. To see this, we plot the adjust Roll measure in Figure 14, and the two measures are remarkably close. Therefore the classic Roll measure will underestimate the bid-ask spread when the order flows display positive autocorrelations.[17] The ReMeDI approach, while being flexible on the autocorrelations patterns, provides an easily implementable and robust measure.

We turn to the ABAS. ABAS measures the average bid-ask spread — an appealing measure to financial regulators, dealers, or investors who execute a large number of transactions, or any party interested in measuring the overall bid-ask spread in a period of time (could be just several seconds). We compare it to the other two measures in Figure 15. The ABAS_n returns persistently larger estimates than the other two measures after incorporating the high order (positive) autocorrelations. Specifically, IBAS_n only captures the instantaneous bid-ask spread and fails to account for the persistence in trades.

8.4 Intraday patterns of bid-ask spread

Empirical microstructure literature has documented that there are intraday patterns of bid-ask spread, see, e.g., Chan and Lakonishok (1995), Hasbrouck (1993), Madhavan et al. (1997), McInish and Wood (1992) and Wood et al. (1985). The spreads typically exhibit a U-shape or reverse J-shape. We demonstrate that our nonparametric measures of bid-ask spread on transaction data produce an L-shape: the spreads are higher in the beginning of a trading day but are relatively constant in the remaining trading hours.

To study the intraday patterns of bid-ask spread, we segment each trading day (from 9:30 AM to 4:00 PM) into 5-minutes intervals and calculate various measures for each interval. The estimates are presented in Figure 16. Several observations are immediate. First, the instantaneous spread measure has an L-shape for each stock. The estimates for the first five minutes are twice as large as the estimates in the remaining sessions. However, we do not find any wider (see McInish and Wood (1992)) or narrower (see Chan et al. (1995)) spreads near the end of the trading day. Using the Roll measure, we find similar patterns, though the estimates are biased. The average spread measure, however, has a slightly different shape — it drops sharply after the first 5 minutes and reverts slowly in the first hour. After that the average spread for INTC remains very constant, but the average spread for KO still exhibits a slightly positive trend.

The spreads are related to market activities. Various measures of market activities display intraday patterns, see Jain and Joh (1988) and McInish and Wood (1990) for trading volume and McInish and Wood (1992) for the number of transactions. To study the interplay between

[17] Based on an AR(1) model of bid-ask spread, Choi et al. (1988) also find that Roll's is downward biased when the spreads are positively autocorrelated.

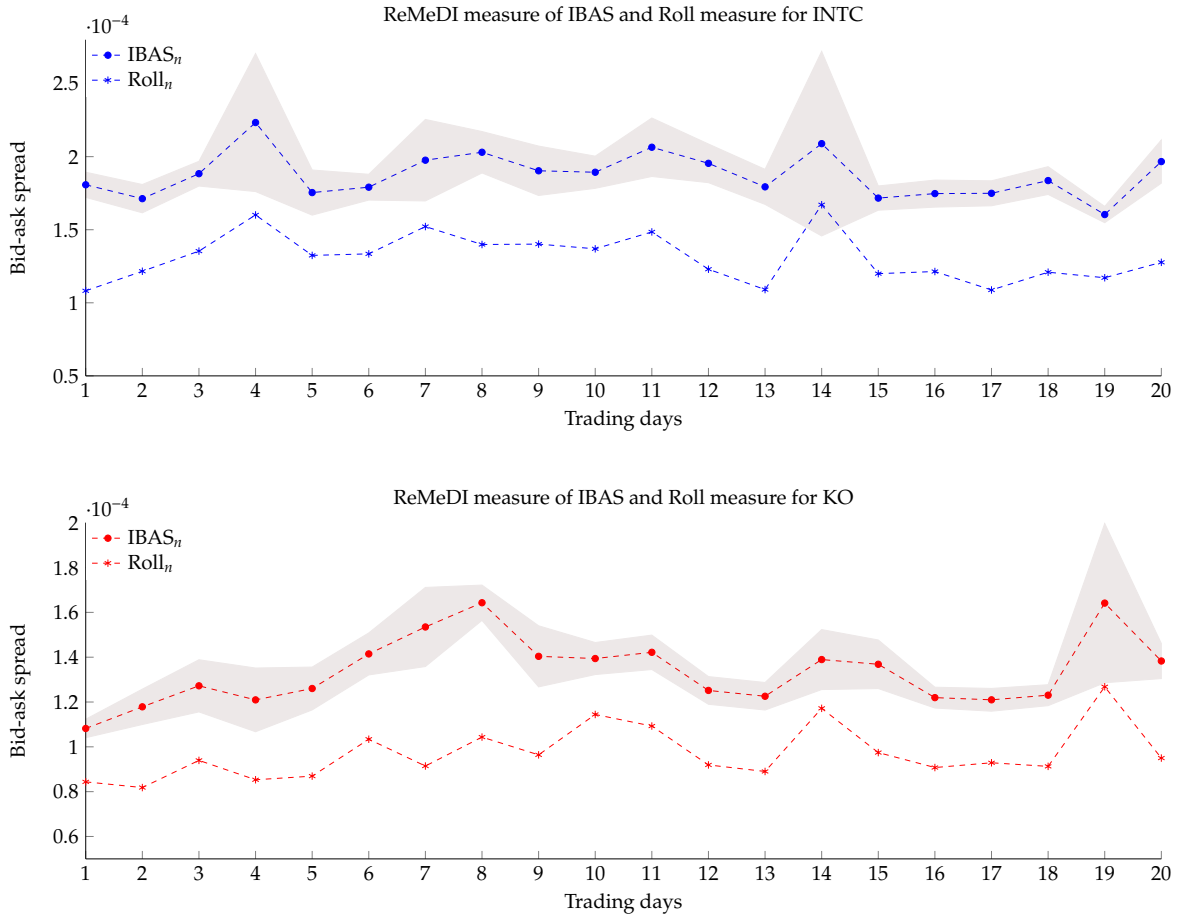

Figure 13: ReMeDI estimates of the instantaneous bid-ask spread (IBAS) and Roll measure for INTC (top panel) and KO (bottom panel) over each trading day in February, 2016, using transaction data from 9:35 AM to 4:00 PM. IBAS$_n$ is constructed in (32) (or (36)), the Roll measure in (44). The shaded area constitutes the 95% confidence intervals of the ReMeDI estimator constructed from Theorem 5.2. The tuning parameters are $k_n = 10, i_n = 6$.

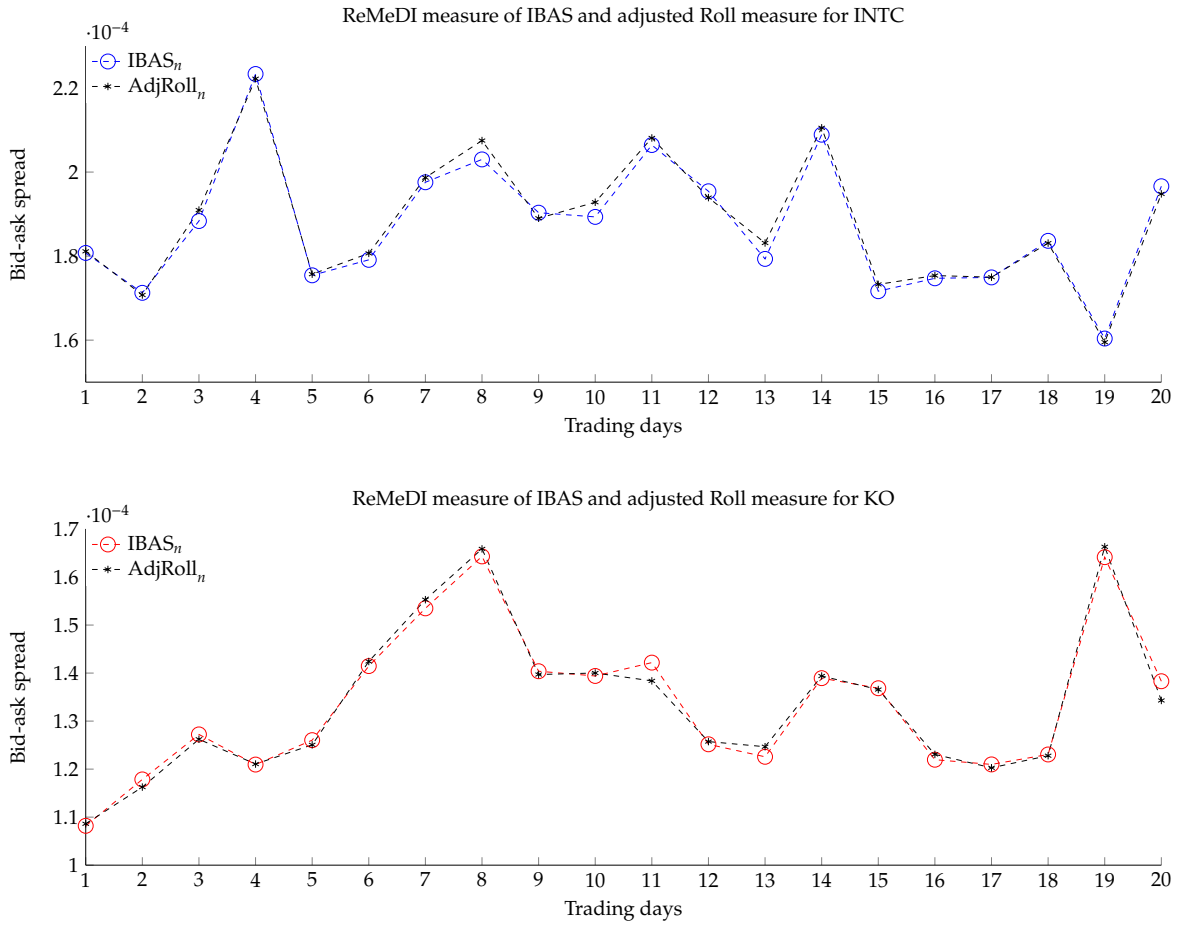

Figure 14: ReMeDI estimates of the instantaneous bid-ask spread (IBAS) and Roll measure for INTC (top panel) and KO (bottom panel) over each trading day in February, 2016, using transaction data from 9:35 AM to 4:00 PM. IBAS$_n$ is constructed in (32) (or (36)), the the adjusted Roll measure in (47), $k_n = 10$.

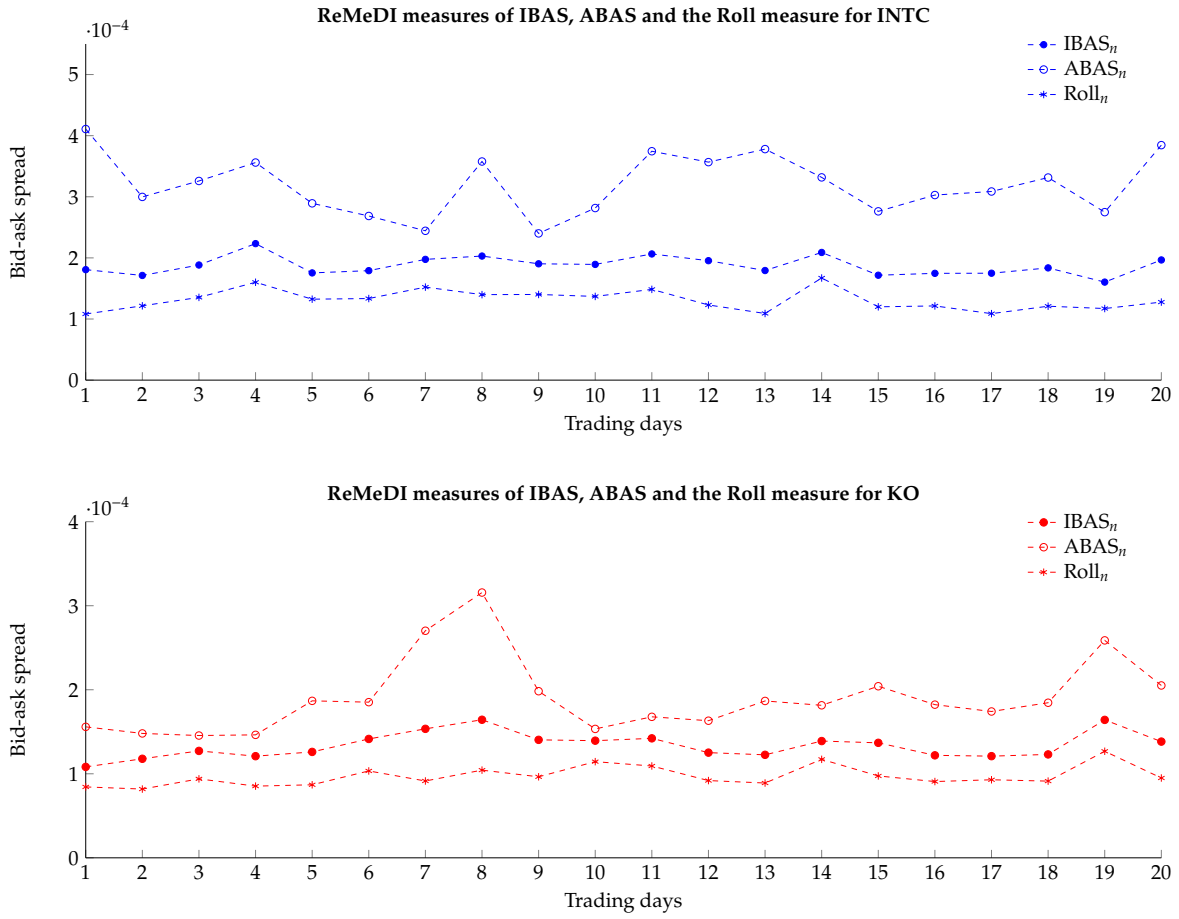

Figure 15: ReMeDI estimates of the instantaneous bid-ask spread (IBAS), average bid-ask spread (ABAS) and Roll measure for INTC (top panel) and KO (bottom panel) over each trading day in February, 2016, using transaction data from 9:35 AM to 4:00 PM. IBAS$_n$ is constructed in (32) (or (36)); ABAS$_n$ is constructed in (33) (or (42)); the Roll measure is constructed in (44). The tuning parameters are $k_n = 10, \ell_n = 6$.

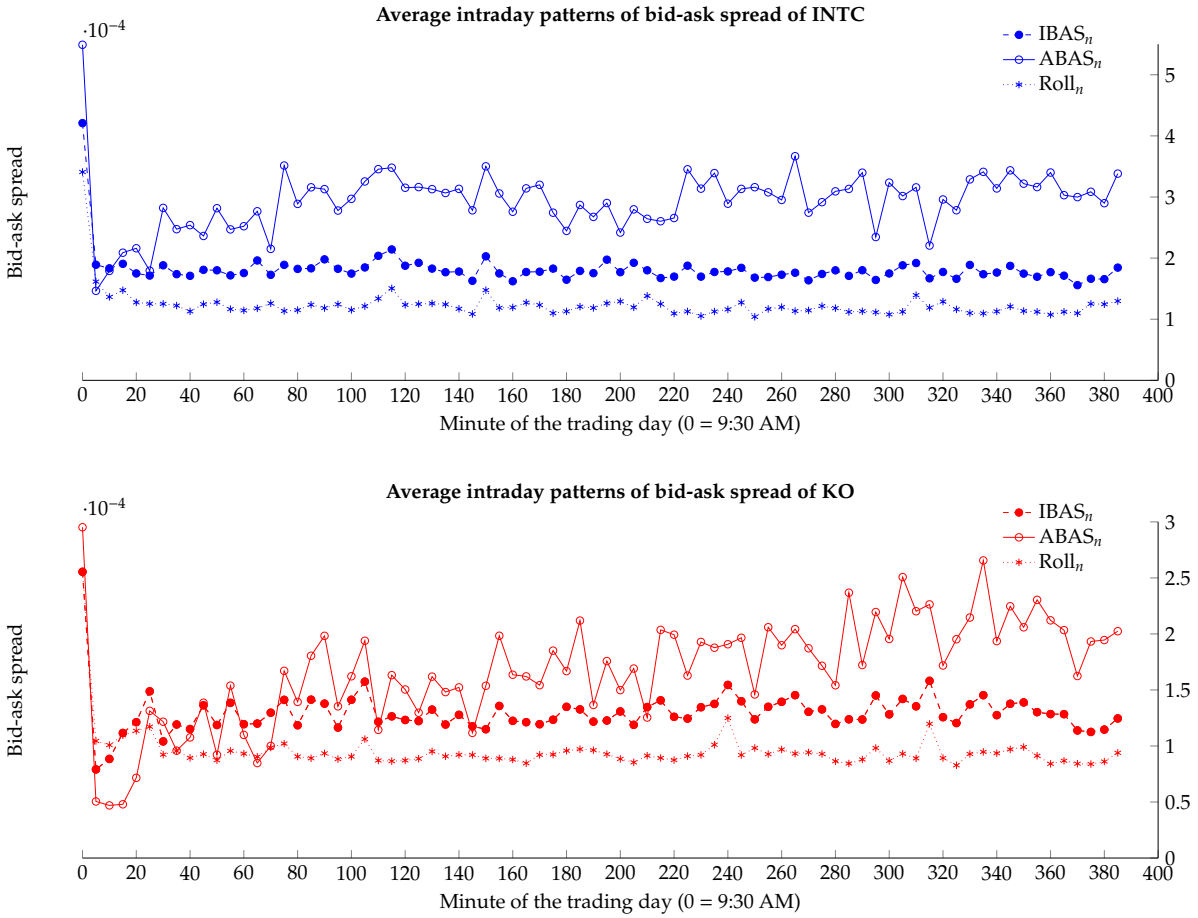

Figure 16: Average intraday patterns of intraday spreads for INTC (top panel) and KO (bottom panel) in February, 2016. The spreads are estimated in 5 minutes local windows on each trading day starting from 9:30 AM to 4:00 PM. IBAS$_n$ is constructed in (32) (or (36)); ABAS$_n$ is constructed in (33) (or (42)); the Roll measure is constructed in (44). The tuning parameters are $k_n = 10, \ell_n = 6$.

intraday spreads and market activities, we report the statistics of the trading volume and number of transactions. The statistics are calculated over 5 minutes time intervals from 9:30 AM to 4:00 PM for each trading day, and the averages of the 20 trading days are plotted in Figure 17. Interestingly, both measures of market activities have U-shapes (see the left and right panels of Figure 17), indicating elevated trading at the beginning and end of the trading day. However, the U-shapes are asymmetric: measured by trading volume, the market is more active in the beginning; in terms of transactions, the market is more active in the close. Thus on average, we would expect large numbers of shares are traded in *each transaction* when the market starts. The middle panel of Figure 17 shows that the *large averages* are contributed by extremely large transactions — the standard deviations of the trading volume in the first five minutes are exceedingly high — trading volumes thereafter, including the end, are quite "smooth". The standard deviations of trading volume display similar L-shape patterns to the spreads. We therefore ascribe the large spreads to extremely large trades, resonating with early findings in Lin et al. (1995) that order processing costs components of bid-ask spread increase in trade sizes for the largest trades.

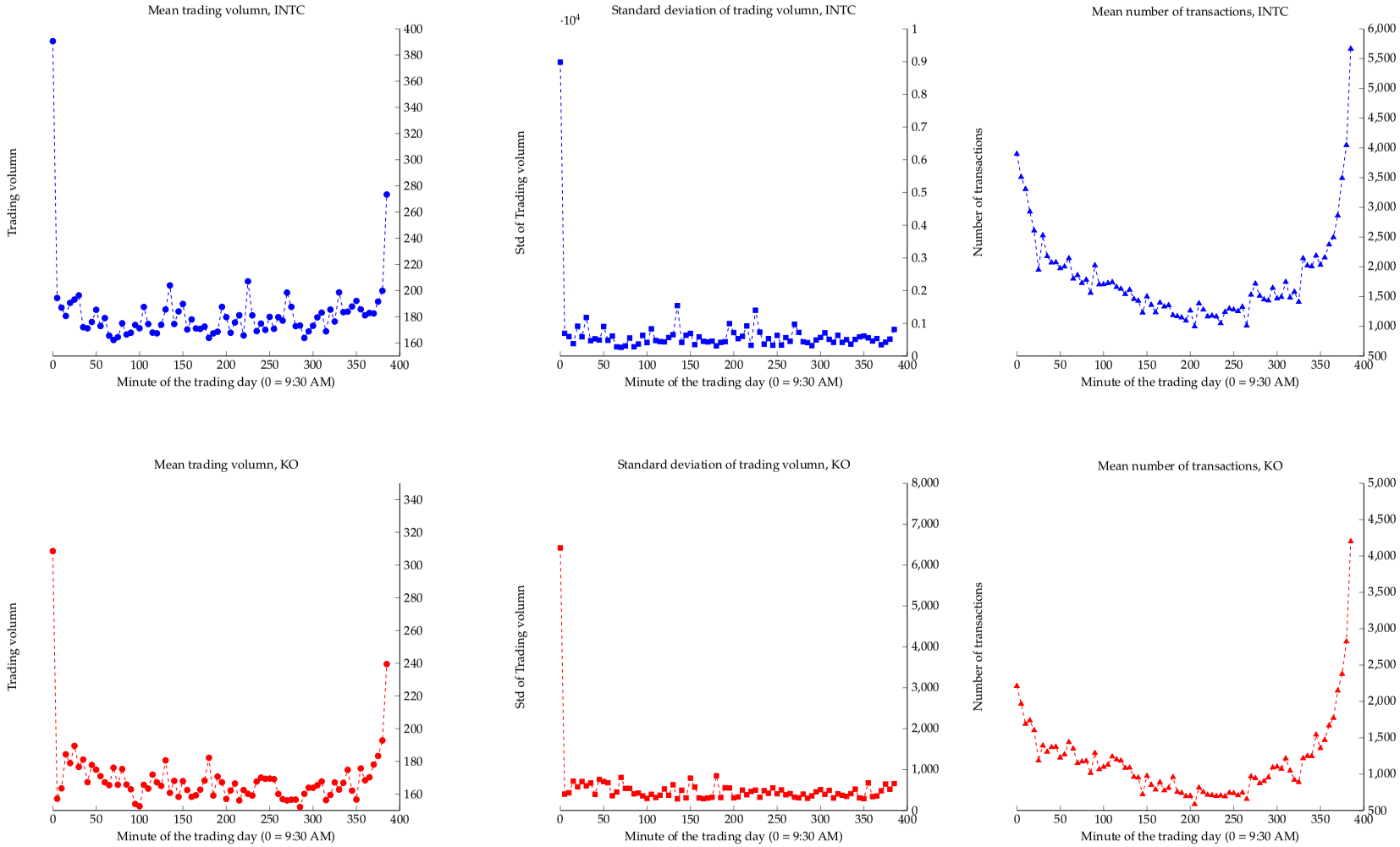

Figure 17: Intraday patterns of mean trading volume (left panel), standard deviation of trading volume (middle panel) and number of transactions (right panel) for INTC (top panel) and KO (bottom panel) in February, 2016. For each of the 20 trading days, the statistics (mean trading volume, standard deviation of trading volume and number of transactions) are calculated over each 5-minutes windows from 9:30 AM to 4:00 PM. The averages of the 20 trading days are plotted.

9 Conclusion

This paper introduces a new econometric method to separate a stationary component from a semimartingale process. The method is robust to data frequencies, model specifications. We also derive the rigorous inferential theory. Based on the proposed estimators, we develop two measures of the bid-ask spread as proxies of market liquidity.

A Some auxiliary results

In the sequel, C denotes a constant that may change from line to line and even within one line. When it depends on some parameters par, we use C_{par}.

Let's first state some classic estimates for Itô semimartingales, one is referred to Jacod and Protter (2011) for details. Let V be any Itô semimartingale satisfying Assumption 3.2. Then for any $s \leq t$, we have for any $r \geq 2$,

$$\mathbb{E}\left(|V_t - V_s|^r \,|\mathcal{F}_s\right) \leq C_r(t-s); \tag{53}$$

We also have

$$\left|\mathbb{E}\left(V_t - V_s \,|\mathcal{F}_s\right)\right| \leq t - s; \tag{54}$$

$$\left|\mathbb{E}\left(\int_{(i-1)\Delta_n}^{i\Delta_n} \left(V_s - V_{i-1}^n\right) \mathrm{d}s \,\middle|\mathcal{F}_{(i-1)\Delta_n}\right)\right| \leq C\Delta_n^2; \tag{55}$$

$$\mathbb{E}\left(\left|\int_{(i-1)\Delta_n}^{i\Delta_n} \left(V_s - V_{i-1}^n\right) \mathrm{d}s\right|^r \,\middle|\mathcal{F}_{(i-1)\Delta_n}\right) \leq C_r\Delta_n^{r+1}. \tag{56}$$

We denote $\mathcal{F}_i^n = \mathcal{F}_{i\Delta_n}$. $\mathcal{G}_i = \sigma(\varepsilon_j : j \leq i), \mathcal{G}^i = \sigma(\varepsilon_j : j \geq i)$; and we introduce $\mathcal{H}_i^n = \mathcal{F}_i^n \otimes \mathcal{G}_i$. The following lemma states an important property of *strongly mixing sequence*, and it will be constantly employed in the proofs.

Lemma A.1. *Let ξ, ξ' be two variables such that ξ is $\mathcal{G}_i$-measurable and ξ' is $\mathcal{G}^{i+k}$-measurable. Assume ξ, ξ' have bounded moments of all orders. Then we have*

$$\left|\mathbb{E}(\xi\xi') - \mathbb{E}(\xi)\mathbb{E}(\xi')\right| \leq Ck^{-v/2}. \tag{57}$$

Proof. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left|\mathbb{E}\left(\left(\xi - \mathbb{E}(\xi)\right)\left(\xi' - \mathbb{E}(\xi')\right)\right)\right| &= \left|\mathbb{E}\left(\left(\xi - \mathbb{E}(\xi)\right)\mathbb{E}\left(\xi' - \mathbb{E}(\xi')\,|\mathcal{G}_i\right)\right)\right| \\ &\leq \sqrt{\mathbb{E}\Big(\left(\xi - \mathbb{E}(\xi)\right)^2\Big)\mathbb{E}\Big(\left(\mathbb{E}\left(\xi' - \mathbb{E}(\xi')\,|\mathcal{G}_i\right)\right)^2\Big)}.\end{aligned}$$

Since ξ' has bounded moments of all orders, Lemma VIII 3.102 of Jacod and Shiryaev (2003) implies $\mathbb{E}\Big(\left(\mathbb{E}\left(\xi' - \mathbb{E}(\xi')\,|\mathcal{G}_i\right)\right)^2\Big) \leq Ck^v$. Now the result follows since $\mathbb{E}\Big(\left(\xi - \mathbb{E}(\xi)\right)^2\Big)$ is bounded. □

B Proof of Theorem 3.1

Proof of Theorem 3.1. We prove the result for $\mathrm{ReMeDI}(Y; j, p)_n^{\mathrm{FF}}$, the consistency of $\mathrm{ReMeDI}(Y; j)_n^{\mathrm{FF}}$ can be proved similarly. For any process V and any $i, j, p \in \mathbb{N}^*$, let

$$\widehat{V}(j,p)_i := -\Delta_{i+j+p}^{k_n} V \Delta_{i+j-2k_n}^{2k_n} V \Delta_{i-3k_n}^{3k_n} V, \quad r(\varepsilon; j, p)_n := \mathbb{E}(\widehat{\varepsilon}(j,p)_i), \quad r(j,p) := \mathbb{E}\Big(\varepsilon_0 \varepsilon_j \varepsilon_{j+p}\Big).$$

Note that

$$\begin{aligned} r(\varepsilon;j,p)_n - r(j,p) =& -\mathbb{E}\Big(\varepsilon_{i+j+p}\varepsilon_{i+j}\varepsilon_{i-3k_n}\Big) - \mathbb{E}\Big(\varepsilon_{i+j+p}\varepsilon_{i+j-2k_n}\varepsilon_i\Big) + \mathbb{E}\Big(\varepsilon_{i+j+p}\varepsilon_{i+j-2k_n}\varepsilon_{i-3k_n}\Big) \\ & - \mathbb{E}\Big(\varepsilon_{i+j+p+k_n}\varepsilon_{i+j}\varepsilon_i\Big) + \mathbb{E}\Big(\varepsilon_{i+j+p+k_n}\varepsilon_{i+j}\varepsilon_{i-3k_n}\Big) - \mathbb{E}\Big(\varepsilon_{i+j+p+k_n}\varepsilon_{i+j-2k_n}\varepsilon_i\Big) \\ & + \mathbb{E}\Big(\varepsilon_{i+j+p+k_n}\varepsilon_{i+j-2k_n}\varepsilon_i\Big). \end{aligned}$$

Apply Lemma A.1, we can show the absolute value of each term on the RHS is bounded by $C/k_n^{v/2}$. We thus have

$$\Big|r(\varepsilon;j,p)_n - r(j,p)\Big| \le \frac{C}{k_n^{v/2}}. \tag{58}$$

For any $k > 5k_n + j + p$, by the independence of X and ε, the martingale property of X and the fact that X has bounded fourth moments and Lemma A.1, we have

$$\begin{aligned} & \Big|\mathbb{E}\Big(\Big(\widehat{Y}(j,p)_i - \widehat{\varepsilon}(j,p)_i\Big)\big(\widehat{\varepsilon}(j,p)_{i+k} - r(\varepsilon;j,p)_n\big)\Big)\Big| \\ =& \Big|\mathbb{E}\Big(\Delta_{i+j+p}^{k_n}\varepsilon\,\big(\widehat{\varepsilon}(j,p)_{i+k} - r(\varepsilon;j,p)_n\big)\Big)\mathbb{E}\Big(\Delta_{i+j-2k_n}^{2k_n}X\Delta_{i-3k_n}^{3k_n}X\Big)\Big| \\ \le& C\Big|\mathbb{E}\Big(\Delta_{i+j+p}^{k_n}\varepsilon\,\big(\widehat{\varepsilon}(j,p)_{i+k} - r(\varepsilon;j,p)_n\big)\Big)\Big| \le C(k-5k_n-j-p)^{-v/2}. \end{aligned}$$

Similar result holds for $\Big|\mathbb{E}\Big(\big(\widehat{\varepsilon}(j,p)_i - r(\varepsilon;j,p)_n\big)\Big(\widehat{Y}(j,p)_{i+k} - \widehat{\varepsilon}(j,p)_{i+k}\Big)\Big)\Big|$, and a direct application of Lemma A.1 yields $\Big|\mathbb{E}\big(\big(\widehat{\varepsilon}(j,p)_i - r(\varepsilon;j,p)_n\big)\big(\widehat{\varepsilon}(j,p)_{i+k} - r(\varepsilon;j,p)_n\big)\big)\Big| \le C(k-5k_n-j-p)^{-v/2}$. Thus we have

$$\begin{aligned} \Big|\mathbb{E}\big(\big(\widehat{\varepsilon}(j,p)_i - r(\varepsilon;j,p)_n\big)\big(\widehat{\varepsilon}(j,p)_{i+k} - r(\varepsilon;j,p)_n\big)\big)\Big| &+ \Big|\mathbb{E}\Big(\Big(\widehat{Y}(j,p)_i - \widehat{\varepsilon}(j,p)_i\Big)\big(\widehat{\varepsilon}(j,p)_{i+k} - r(\varepsilon;j,p)_n\big)\Big)\Big| \\ &+ \Big|\mathbb{E}\Big(\big(\widehat{\varepsilon}(j,p)_i - r(\varepsilon;j,p)_n\big)\Big(\widehat{Y}(j,p)_{i+k} - \widehat{\varepsilon}(j,p)_{i+k}\Big)\Big)\Big| \\ &\le C(k-5k_n-j-p)^{-v/2}. \end{aligned} \tag{59}$$

Next, we show

$$\Big|\mathbb{E}\Big(\Big(\widehat{Y}(j,p)_i - \widehat{\varepsilon}(j,p)_i\Big)\Big(\widehat{Y}(j,p)_{i+k} - \widehat{\varepsilon}(j,p)_{i+k}\Big)\Big)\Big| \le C(k-5k_n-j-p)^{-v/2}. \tag{60}$$

We get 49 terms after expanding $\Big(\widehat{Y}(j,p)_i - \widehat{\varepsilon}(j,p)_i\Big)\Big(\widehat{Y}(j,p)_{i+k} - \widehat{\varepsilon}(j,p)_{i+k}\Big)$. We divide the 49 terms into two categories. First, if the term has exact one of $\Delta_{i+k+j+p}^{k_n}X, \Delta_{i+k+j-2k_n}^{2k_n}X, \Delta_{i+k-3k_n}^{3k_n}X$, then its expectation is zero by the martingale property of X. Second, if the term has at most one of $\Delta_{i+k+j+p}^{k_n}\varepsilon, \Delta_{i+k+j-2k_n}^{2k_n}\varepsilon, \Delta_{i+k-3k_n}^{3k_n}\varepsilon$, then its expectation is either zero (by the independence of X and ε, the martingale property of X, and the fact $\mathbb{E}(\varepsilon)=0$) or bounded by $C(k-5k_n-j-p)^{-v/2}$ (by the independence of X and ε and Lemma A.1). This proves (60). Combined with (59), we get

$$\Big|\mathbb{E}\Big(\Big(\widehat{Y}(j,p)_i - r(\varepsilon;j,p)_n\Big)\Big(\widehat{Y}(j,p)_{i+k} - r(\varepsilon;j,p)_n\Big)\Big)\Big| \le C(k-5k_n-j-p)^{-v/2},$$

which further leads to (recall $v > 2$)

$$\mathbb{E}\left(\left(\sum_{i=3k_n}^{n-k_n-j}\left(\widehat{Y}(j,p)_i - r(\varepsilon;j,p)_n\right)\right)^2\right) \le Cnk_n. \tag{61}$$

In view of (58) and (61), we have

$$\mathbb{E}\left(n^{-2}\left(\sum_{i=3k_n}^{n-k_n-j}\left(\widehat{Y}(j,p)_i - r(j,p)\right)\right)^2\right) \le \max\{k_n^{-v}, k_n/n\} \to 0. \tag{62}$$

This proves $\mathrm{ReMeDI}(Y;j,p)_n^{\mathrm{FF}} \xrightarrow{\mathbb{P}} r(j,p)$. □

C Proof of Theorem 3.2

Proof. We prove (18), the proofs of (20), (21) and (45) are easier and can be proved in a similar way. For a process V, let $\widehat{V}(j,p,q)_i^n = -\Delta_{i-4k_n}^{n,4k_n} V \Delta_{i+j-3k_n}^{n,3k_n} V \Delta_{i+j+p-2k_n}^{n,2k_n} V \Delta_{i+j+p+q}^{n,k_n} V$. We obtain the following using similar arguments to obtain (58):

$$\left|\mathbb{E}\left(\widehat{\varepsilon}(j,p,q)_i^n\right) - \mathbb{E}\left(\varepsilon_i\varepsilon_{i+j}\varepsilon_{i+j+p}\varepsilon_{i+j+p+q}\right)\right| \le \frac{C}{k_n^{v/2}}. \tag{63}$$

We denote $\ell_0 = 0, \ell_1 = j, \ell_2 = j+p, \ell_3 = j+p+q$, and let

$$\delta_{i,k}^n = \begin{cases} X_{i+\ell_k}^n - X_{i+\ell_k-(4-k)k_n}^n & k = 0,1,2; \\ X_{i+\ell_k}^n - X_{i+\ell_k+k_n}^n & k = 3. \end{cases} \quad \tilde{\delta}_{i,k}^n = \begin{cases} \varepsilon_{i+\ell_k}^n - \varepsilon_{i+\ell_k-(4-k)k_n}^n & k = 0,1,2; \\ \varepsilon_{i+\ell_k}^n - \varepsilon_{i+\ell_k+k_n}^n & k = 3. \end{cases}$$

Note that

$$\widehat{Y}(j,p,q)_i^n = \prod_{k=0}^{3}\left(\delta_{i,k} + \tilde{\delta}_{i,k}\right); \quad \widehat{\varepsilon}(j,p,q)_i^n = \prod_{k=0}^{3}\tilde{\delta}_{i,k}.$$

And

$$\widehat{Y}(j,p,q)_i^n - \widehat{\varepsilon}(j,p,q)_i^n = \sum_{(Q,Q^c)\subset\mathcal{Q}_4}\prod_{k\in Q}\delta_{i,k}\prod_{k'\in Q^c}\tilde{\delta}_{i,k'}, \tag{64}$$

where $\mathcal{Q}_4 = \{(Q,Q^c) : Q\cap Q^c = \emptyset, Q\cup Q^c = \{0,1,2,3\}, Q\neq\emptyset\}$. By (53), and the fact that all moments of noise exist, we have for any $r \ge 2, (Q,Q^c) \in \mathcal{Q}_4$,

$$\mathbb{E}\left(\left|\delta_{i,k}^n\right|^r\right) \le Ck_n\Delta_n, \quad \mathbb{E}\left(\left|\prod_{k'\in Q^c}\tilde{\delta}_{i,k}^n\right|^r\right) \le C. \tag{65}$$

Given $(Q,Q^c) \in \mathcal{Q}_4$, let $l = |Q|$, $l \ge 1$ since $Q \neq \emptyset$. Apply the (generalized) Hölder's inequality

with exponents $(\underbrace{2l,\dots,2l}_{l},2)$, we have

$$\mathbb{E}\left(\left|\prod_{k\in Q}\delta^n_{i,k}\prod_{k'\in Q^c}\tilde{\delta}^n_{i,k'}\right|\right)\le\prod_{k\in Q}\left(\mathbb{E}\left(\left|\delta^n_{i,k}\right|^{2l}\right)\right)^{\frac{1}{2l}}\sqrt{\mathbb{E}\left(\left|\prod_{k'\in Q^c}\tilde{\delta}^n_{i,k'}\right|^2\right)}\overset{(65)}{\le}C\sqrt{k_n\Delta_n}. \tag{66}$$

Now it follows from (63), (64) and (66), and the facts that $k_n\to\infty$ and $k_n\Delta_n\to 0$:

$$\begin{aligned}&\mathbb{E}\left(\left|\mathrm{ReMeDI}(Y;j,p,q)^{\mathrm{HF}}_n-\mathbb{E}\left(\varepsilon_i\varepsilon_{i+j}\varepsilon_{i+j+p}\varepsilon_{i+j+p+q}\right)\right|\right)\\&\le\sum_{i=4k_n}^{N^n_t-k_n-j-p-q}\frac{\mathbb{E}\left(\left|\widehat{Y}(j,p,q)^n_i-\widehat{\varepsilon}(j,p,q)^n_i\right|\right)+\mathbb{E}\left(\left|\widehat{\varepsilon}(j,p,q)^n_i-\mathbb{E}\left(\varepsilon_i\varepsilon_{i+j}\varepsilon_{i+j+p}\varepsilon_{i+j+p+q}\right)\right|\right)}{N^n_t-5k_n-j-p-q+1}\to 0.\end{aligned}$$

The proof is finished. □

D Proof of Theorem 3.3

For any process V, let $\widehat{V}^n_i=-\Delta^{n,2k_n}_{i-2k_n}V\Delta^{n,k_n}_{i+j}V$.

Lemma D.1. *Let ε satisfy Assumption 2.1. For any $k_n\to\infty$, we have*

$$r^n_j:=\mathbb{E}\left(\widehat{\varepsilon}^n_i\right)\to r_j. \tag{67}$$

For any $k\in\mathbb{Z}$,

$$\left|\mathbb{E}\left(\left(\widehat{\varepsilon}^n_i-r^n_j\right)\left(\widehat{\varepsilon}^n_{i+k}-r^n_j\right)\right)-3r^2_k-\mathbb{E}\left(\left(\varepsilon_i\varepsilon_{i+j}-r_j\right)\left(\varepsilon_{i+k}\varepsilon_{i+k+j}-r_j\right)\right)\right|\le\frac{C}{k^{v/2}_n}. \tag{68}$$

Proof. It is immediate to get

$$\left|r^n_j-r_j\right|\le\frac{C}{k^{v/2}_n} \tag{69}$$

after an application of Lemma A.1.To prove (68), we consider the following scenarios:

(a) We show for any $k\in\mathbb{Z}$,

$$\begin{cases}\left|\mathbb{E}\left(\varepsilon^n_i\varepsilon^n_{i+k}\varepsilon^n_{i+j+k_n}\varepsilon^n_{i+j+k_n+k}\right)-r^2_k\right|\le\frac{C}{k^{v/2}_n};\\\left|\mathbb{E}\left(\varepsilon^n_{i-2k_n}\varepsilon^n_{i-2k_n+k}\varepsilon^n_{i+j}\varepsilon^n_{i+j+k}\right)-r^2_k\right|\le\frac{C}{k^{v/2}_n};\\\left|\mathbb{E}\left(\varepsilon^n_{i-2k_n}\varepsilon^n_{i-2k_n+k}\varepsilon^n_{i+j+k_n}\varepsilon^n_{i+j+k_n+k}\right)-r^2_k\right|\le\frac{C}{k^{v/2}_n}.\end{cases} \tag{70}$$

(1) To prove the first estimate in (70), we first consider $|k|\le\lfloor k_n/2\rfloor$ so that

$$\min(i+j+k_n+k,i+j+k_n)-\max(i,i+k)\ge k_n/2.$$

Then apply Lemma A.1, we have

$$\left|\mathbb{E}\Big(\big(\varepsilon_i^n\varepsilon_{i+k}^n - r_k\big)\big(\varepsilon_{i+j+k_n}^n\varepsilon_{i+k+j+k_n}^n - r_k\big)\Big)\right| \le \frac{C}{k_n^{v/2}}.$$

When $|k| > \lfloor k_n/2 \rfloor$, we first note $r_k^2 \le C/k_n^v$ by (3), whereas

$$\begin{aligned}
&\left|\mathbb{E}\Big(\varepsilon_i^n\varepsilon_{i+j+k_n}^n\varepsilon_{i+k}^n\varepsilon_{i+k+j+k_n}^n\Big)\right| \\
\le&\begin{cases}
\sqrt{\mathbb{E}\left(\left(\varepsilon_i^n\varepsilon_{i+k}^n\varepsilon_{i+j+k_n}^n\right)^2\right)\mathbb{E}\left(\left(\mathbb{E}\left(\varepsilon_{i+j+k_n+k}^n\,\middle|\mathcal{G}_{\max(i+j+k_n,i+k)}\right)\right)^2\right)}, & k > \lfloor k_n/2\rfloor \\
\sqrt{\mathbb{E}\left(\left(\varepsilon_i^n\varepsilon_{i+j+k_n+k}^n\varepsilon_{i+j+k_n}^n\right)^2\right)\mathbb{E}\left(\left(\mathbb{E}\left(\varepsilon_{i+k}^n\,\middle|\mathcal{G}^{\min(i,i+k+j+k_n)}\right)\right)^2\right)}, & k < -\lfloor k_n/2\rfloor
\end{cases} \\
\le&\frac{C}{k_n^{v/2}}.
\end{aligned}$$

(2) For the second estimate in (70), we first assume $|k| \le k_n$ thus

$$\min(i+j, i+j+k) - \max(i-2k_n, i-2k_n+k) \ge k_n$$

so that we have

$$\begin{aligned}
&\left|\mathbb{E}\Big(\big(\varepsilon_{i-2k_n}^n\varepsilon_{i+k-2k_n}^n - r_k\big)\big(\varepsilon_{i+j}^n\varepsilon_{i+j+k}^n - r_k\big)\Big)\right| \\
\le&\sqrt{\mathbb{E}\Big(\varepsilon_{i-2k_n}^n\varepsilon_{i+k-2k_n}^n - r_k\Big)^2\left(\mathbb{E}\left(\mathbb{E}\left(\varepsilon_{i+j}^n\varepsilon_{i+j+k}^n - r_k\,\middle|\mathcal{G}_{\max(i-2k_n,i+k-2k_n)}\right)\right)^2\right)} \le \frac{C}{k_n^{v/2}}.
\end{aligned}$$

For $|k| > k_n$, we have $r_k^2 \le C/k_n^v$ and

$$\begin{aligned}
&\left|\mathbb{E}\Big(\varepsilon_{i-2k_n}^n\varepsilon_{i-2k_n+k}^n\varepsilon_{i+j}^n\varepsilon_{i+k+j}^n\Big)\right| \\
\le&\begin{cases}
\sqrt{\mathbb{E}\left(\left(\varepsilon_{i-2k_n}^n\varepsilon_{i-2k_n+k}^n\varepsilon_{i+j}^n\right)^2\right)\mathbb{E}\left(\left(\mathbb{E}\left(\varepsilon_{i+j+k}^n\,\middle|\mathcal{G}_{\max(i+j,i+k-2k_n)}\right)\right)^2\right)}, & k > k_n \\
\sqrt{\mathbb{E}\left(\left(\varepsilon_{i-2k_n}^n\varepsilon_{i+j}^n\varepsilon_{i+j+k}^n\right)^2\right)\mathbb{E}\left(\left(\mathbb{E}\left(\varepsilon_{i-2k_n+k}^n\,\middle|\mathcal{G}^{\min(i-2k_n,i+k+j)}\right)\right)^2\right)}, & k < -k_n
\end{cases} \\
\le&\frac{C}{k_n^{v/2}}.
\end{aligned}$$

(3) The last estimate in (70) can be proved similarly. For $|k| \le k_n$ thus

$$\min(i+j+k_n, i+j+k+k_n) - \max(i-2k_n, i-2k_n+k) \ge k_n$$

so that we have

$$\begin{aligned}&\left|\mathbb{E}\Big(\Big(\varepsilon^n_{i-2k_n}\varepsilon^n_{i+k-2k_n}-r_k\Big)\Big(\varepsilon^n_{i+j+k_n}\varepsilon^n_{i+j+k_n+k}-r_k\Big)\Big)\right|\\ \le&\sqrt{\mathbb{E}\Big(\varepsilon^n_{i-2k_n}\varepsilon^n_{i+k-2k_n}-r_k\Big)^2\left(\mathbb{E}\Big(\mathbb{E}\left(\varepsilon^n_{i+j+k_n}\varepsilon^n_{i+j+k+k_n}-r_k\,\middle|\mathcal{G}_{\max(i-2k_n,i-2k_n+k)}\right)\Big)^2\right)}\le\frac{C}{k_n^{v/2}}.\end{aligned}$$

For $|k|>k_n$, we have $r_k^2\le C/k_n^v$ and

$$\begin{aligned}&\left|\mathbb{E}\Big(\varepsilon^n_{i-2k_n}\varepsilon^n_{i-2k_n+k}\varepsilon^n_{i+j+k_n}\varepsilon^n_{i+k+j+k_n}\Big)\right|\\ \le&\begin{cases}\sqrt{\mathbb{E}\left(\left(\varepsilon^n_{i-2k_n}\varepsilon^n_{i-2k_n+k}\varepsilon^n_{i+j+k_n}\right)^2\right)\mathbb{E}\left(\left(\mathbb{E}\left(\varepsilon^n_{i+j+k+k_n}\,\middle|\mathcal{G}_{\max(i+j+k_n,i+k-2k_n)}\right)\right)^2\right)}, & k>k_n\\ \sqrt{\mathbb{E}\left(\left(\varepsilon^n_{i-2k_n}\varepsilon^n_{i+j+k_n}\varepsilon^n_{i+j+k+k_n}\right)^2\right)\mathbb{E}\left(\left(\mathbb{E}\left(\varepsilon^n_{i-2k_n+k}\,\middle|\mathcal{G}^{\min(i-2k_n,i+k+j+k_n)}\right)\right)^2\right)}, & k<-k_n\end{cases}\\ \le&\frac{C}{k_n^{v/2}}.\end{aligned}$$

This finishes the proof of (70).

(b) Now we show the remaining 13 terms in $\mathbb{E}\left(\widehat{\varepsilon}^n_{i+k}\widehat{\varepsilon}^n_i\right)$ are bounded by $C/k_n^{v/2}$ except $\varepsilon^n_i\varepsilon^n_{i+j}\varepsilon^n_{i+k}\varepsilon^n_{i+k+j}$. The approach is still to apply Lemma A.1 to "separate" each term. Therefore we only need to show the maximal or minimal index is at least Ck_n larger or smaller than the remaining indices.

(1) For $\varepsilon^n_i\varepsilon^n_{i+j}\varepsilon^n_{i+k-2k_n}\varepsilon^n_{i+k+j+k_n}$, we have

$$\begin{cases}\min(i,i+j,i+k+j+k_n)-(i+k-2k_n)\ge 2k_n, & k<0\\ (i+k+j+k_n)-\max(i+j,i,i+k-2k_n)\ge k_n. & k\ge 0\end{cases}$$

(2) For $\varepsilon^n_i\varepsilon^n_{i+j}\varepsilon^n_{i+k-2k_n}\varepsilon^n_{i+k+j}$, we have

$$\begin{cases}\min(i,i+j,i+k+j)-(i+k-2k_n)\ge\lfloor 3k_n/2\rfloor, & k<\lfloor k_n/2\rfloor\\ (i+k+j)-\max(i+j,i,i+k-2k_n)\ge\lfloor k_n/2\rfloor. & k\ge\lfloor k_n/2\rfloor\end{cases}$$

(3) For $\varepsilon^n_i\varepsilon^n_{i+j}\varepsilon^n_{i+k}\varepsilon^n_{i+k+k_n}$, we have

$$\begin{cases}\min(i,i+j,i+k+j)-(i+k)>\lfloor k_n/2\rfloor, & k<-\lfloor k_n/2\rfloor\\ (i+k+k_n)-\max(i,i+j,i+k)\ge\lfloor k_n/2\rfloor-j. & k\ge-\lfloor k_n/2\rfloor\end{cases}$$

(4) For $\varepsilon_i^n \varepsilon_{i+j+k_n}^n \varepsilon_{i+k-2k_n}^n \varepsilon_{i+k+j+k_n}^n$, we have

$$\begin{cases} \min(i, i+j+k_n, i+k+j+k_n) - (i+k-2k_n) \geq k_n, & k < k_n \\ (i+k+j+k_n) - \max(i, i+j+k_n, i+k-2k_n) \geq k_n. & k \geq k_n \end{cases}$$

(5) For $\varepsilon_i^n \varepsilon_{i+j+k_n}^n \varepsilon_{i+k-2k_n}^n \varepsilon_{i+k+j}^n$, we have

$$\begin{cases} \min(i, i+j+k_n, i+k+j) - (i+k-2k_n) \geq \lfloor k_n/2 \rfloor, & k < \lfloor 3k_n/2 \rfloor \\ (i+k+j) - \max(i+j+k_n, i, i+k-2k_n) \geq \lfloor k_n/2 \rfloor. & k \geq \lfloor 3k_n/2 \rfloor \end{cases}$$

(6) For $\varepsilon_i^n \varepsilon_{i+j+k_n}^n \varepsilon_{i+k}^n \varepsilon_{i+k+j}^n$, we have

$$\begin{cases} i+j+k_n - \max(i, i+k+j, i+k) \geq \lfloor k_n/2 \rfloor, & k < \lfloor k_n/2 \rfloor \\ \min(i+j+k_n, i+k, i+k+j) - i \geq \lfloor k_n/2 \rfloor. & k \geq \lfloor k_n/2 \rfloor \end{cases}$$

(7) For $\varepsilon_{i-2k_n}^n \varepsilon_{i+j}^n \varepsilon_{i+k}^n \varepsilon_{i+k+j+k_n}^n$, we have

$$\begin{cases} \min(i+k, i+j, i+k+j+k_n) - (i-2k_n) \geq \lfloor k_n/2 \rfloor, & k \geq \lfloor -3k_n/2 \rfloor \\ (i+j) - \max(i-2k_n, i+k, i+k+j+k_n) \geq \lfloor k_n/2 \rfloor. & k < \lfloor -3k_n/2 \rfloor \end{cases}$$

(8) For $\varepsilon_{i-2k_n}^n \varepsilon_{i+j}^n \varepsilon_{i+k-2k_n}^n \varepsilon_{i+k+j+k_n}^n$, we have

$$\begin{cases} \min(i-2k_n, i+j, i+k+j+k_n) - (i+k-2k_n) \geq \lfloor k_n/2 \rfloor, & k < \lfloor -k_n/2 \rfloor \\ (i+j+k+k_n) - \max(i-2k_n, i+j, i+k-2k_n) \geq \lfloor k_n/2 \rfloor. & k \geq \lfloor -k_n/2 \rfloor \end{cases}$$

(9) For $\varepsilon_{i-2k_n}^n \varepsilon_{i+j+k_n}^n \varepsilon_{i+k}^n \varepsilon_{i+k+j}^n$, we have

$$\begin{cases} \min(i+k, i+j+k, i+j+k_n) - (i-2k_n) \geq 2k_n, & k \geq 0 \\ (i+j+k_n) - \max(i-2k_n, i+k, i+k+j) \geq k_n. & k < 0 \end{cases}$$

(10) For $\varepsilon_{i-2k_n}^n \varepsilon_{i+j+k_n}^n \varepsilon_{i+k}^n \varepsilon_{i+k+j+k_n}^n$, we have

$$\begin{cases} \min(i+k, i+j+k_n, i+j+k+k_n) - (i-2k_n) \geq k_n, & k \geq -k_n \\ (i+j+k_n) - \max(i-2k_n, i+k, i+k+j+k_n) \geq k_n. & k < -k_n \end{cases}$$

(11) For $\varepsilon_{i-2k_n}^n \varepsilon_{i+j+k_n}^n \varepsilon_{i+k-2k_n}^n \varepsilon_{i+k+j}^n$, we have

$$\begin{cases} \min(i+k-2k_n, i+k+j, i+j+k_n) - (i-2k_n) \geq \lfloor k_n/2 \rfloor, & k \geq \lfloor k_n/2 \rfloor \\ (i+j+k_n) - \max(i-2k_n, i+k-2k_n, i+k+j) \geq \lfloor k_n/2 \rfloor. & k < \lfloor k_n/2 \rfloor \end{cases}$$

(12) For $\varepsilon^n_{i-2k_n}\varepsilon^n_{i+j}\varepsilon^n_{i+k}\varepsilon^n_{i+k+j}$, we have

$$\begin{cases} \min(i+k, i+j, i+j+k) - (i-2k_n) \geq k_n, & k \geq -k_n \\ (i+j) - \max(i-2k_n, i+k, i+k+j) \geq k_n. & k < -k_n \end{cases}$$

Now the proof of (68) is complete.

□

Lemma D.2. *Let $v > 2, k_n\Delta_n \to 0$. Then*

$$\Sigma^n_j := \frac{\mathbf{Var}\left(\sum_{i=2k_n}^{N^n_t - k_n - j} \widehat{\varepsilon}^n_i\right)}{N^n_t - 3k_n - j + 1} \to \Sigma_j. \tag{71}$$

Proof. Since $v > 2$, (68) implies

$$\left| \sum_{k=-j-4k_n}^{j+4k_n} \mathbb{E}\Big(\big(\widehat{\varepsilon}^n_i - r^n_j\big)\big(\widehat{\varepsilon}^n_{i+k} - r^n_j\big)\Big) - \sum_{k=-4k_n-j}^{4k_n+j} \Big(3r_k^2 + \mathbb{E}\Big(\big(\varepsilon_0\varepsilon_j - r_j\big)\big(\varepsilon_k\varepsilon_{k+j} - r_j\big)\Big)\Big) \right| \leq \frac{Ck_n}{k_n^{v/2}} \to 0. \tag{72}$$

We also have

$$\sum_{k=4k_n+j}^{\infty} \Big(3r_k^2 + \mathbb{E}\Big(\big(\varepsilon_0\varepsilon_j - r_j\big)\big(\varepsilon_k\varepsilon_{k+j} - r_j\big)\Big)\Big) \leq \frac{C}{k_n^{\frac{v}{2}-1}} \to 0. \tag{73}$$

Hence (72) and (73) imply

$$\sum_{k=-j-4k_n}^{j+4k_n} \mathbb{E}\Big(\big(\widehat{\varepsilon}^n_i - r^n_j\big)\big(\widehat{\varepsilon}^n_{i+k} - r^n_j\big)\Big) \to \Sigma_j. \tag{74}$$

Since

$$\left| \sum_{k=j+4k_n}^{\infty} \mathbb{E}\Big(\big(\widehat{\varepsilon}^n_i - r^n_j\big)\big(\widehat{\varepsilon}^n_{i+k} - r^n_j\big)\Big) \right| \leq \frac{C}{k_n^{\frac{v}{2}-1}} \to 0,$$

(71) now follows. □

Lemma D.3. *Let $(w_n)_n$ be a sequence of integers such that $w_n > 2k_n$. Then for $v > 2$, we have*

$$\mathbb{E}\left(\left(\sum_{i=2k_n}^{w_n} \left(\widehat{\varepsilon}^n_i - r^n_j\right)\right)^2\right) \leq C\left(w_n - 2k_n\right). \tag{75}$$

Proof. It suffices to prove for each $2k_n \leq i \leq w_n - 1$,

$$\mathbb{E}\left(\sum_{k=1}^{w_n - i} \left(\widehat{\varepsilon}^n_i - r^n_j\right)\left(\widehat{\varepsilon}^n_{i+k} - r^n_j\right)\right) \leq C. \tag{76}$$

Note that

$$\left|\mathbb{E}\left(\sum_{k=1}^{w_n-i}\left(\widehat{\varepsilon}_i^n-r_j^n\right)\left(\widehat{\varepsilon}_{i+k}^n-r_j^n\right)\right)\right| \le \left|\mathbb{E}\left(\sum_{k=1}^{(w_n-i)\wedge(3k_n+j)}\left(\widehat{\varepsilon}_i^n-r_j^n\right)\left(\widehat{\varepsilon}_{i+k}^n-r_j^n\right)\right)\right| + \left|\mathbb{E}\left(\sum_{k=1+(3k_n+j)}^{(w_n-i)\vee(3k_n+j)}\left(\widehat{\varepsilon}_i^n-r_j^n\right)\left(\widehat{\varepsilon}_{i+k}^n-r_j^n\right)\right)\right|.$$

But for $k \ge 1+(3k_n+j)$, $\left|\mathbb{E}\left(\left(\widehat{\varepsilon}_i^n-r_j^n\right)\left(\widehat{\varepsilon}_{i+k}^n-r_j^n\right)\right)\right| \le \frac{C}{(k-3k_n-j)^{v/2}}$. Recall that $v>2$, thus it suffices to prove $\sum_{k=1}^{(w_n-i)\wedge(3k_n+j)}\left|\mathbb{E}\left(\left(\widehat{\varepsilon}_i^n-r_j^n\right)\left(\widehat{\varepsilon}_{i+k}^n-r_j^n\right)\right)\right| < C$, which is a simple consequence of (68).

□

Theorem D.1. *Let the noise process ε satisfy Assumption 2.1, k_n, v satisfy*

$$v>6, \quad k_n \asymp \Delta_n^{-\gamma},\ \gamma\in(0,\delta)\,, \text{where } \delta\in\left(\frac{2}{v+4},\frac{1}{5}\right). \tag{77}$$

Then we have the following convergence in distribution:

$$\sqrt{N_t^n}\left(\mathrm{ReMeDI}(\varepsilon;j)_n^{\mathrm{HF}}-r_j^n\right)\xrightarrow{\mathcal{L}}\mathcal{N}\left(0,\Sigma_j\right). \tag{78}$$

In addition if $\gamma>\frac{1}{v}$, we have

$$\sqrt{N_t^n}\left(\mathrm{ReMeDI}(\varepsilon;j)_n^{\mathrm{HF}}-r_j\right)\xrightarrow{\mathcal{L}}\mathcal{N}\left(0,\Sigma_j\right). \tag{79}$$

Proof. We follow the steps to prove central limit theorems of dependent variables, see Ibragimov (1962). Let $(p_n),(q_n)$ be two sequences of integers satisfying

$$p_n\asymp\Delta_n^{-\tau},\tau\in\left(2\delta,\frac{1-\delta}{2}\right);\quad q_n\asymp\Delta_n^{-\ell},\ell\in(\delta,2\delta)\,. \tag{80}$$

Let $v_n=[(N_t^n-k_n-j)/(p_n+q_n)]$. Thus $v_n\asymp\Delta_n^{-(1-\tau)}$. We can verify the following asymptotic relations:

$$p_n^{-1}q_n\to 0; \tag{81}$$
$$q_n^{-1}k_n\to 0; \tag{82}$$
$$k_np_n^2\Delta_n\to 0; \tag{83}$$
$$q_n^{-\frac{v}{2}}v_n\to 0. \tag{84}$$

In particular, the last asymptotic relation holds since

$$1-\tau<1-2\delta<1-\frac{4}{v+4}=\frac{v}{2}\times\frac{2}{v+4}<\frac{v}{2}\delta<\frac{v}{2}\ell.$$

For $1 \le k \le \nu_n - 1$, let

$$\tilde{\eta}_k^n = \sum_{i=k(p_n+q_n)+1}^{k(p_n+q_n)+q_n} \left(\widehat{\varepsilon}_i^n - r_j^n\right); \quad \bar{\eta}_k^n = \sum_{i=k(p_n+q_n)+q_n+1}^{(k+1)(p_n+q_n)} \left(\widehat{\varepsilon}_i^n - r_j^n\right). \tag{85}$$

Recall r_j^n is defined in (67). As we will prove later, the "big" blocks $(\bar{\eta}_k^n)_k$, separated by the "small" blocks $(\tilde{\eta}_k^n)_k$ are approximately independent, while the "small" blocks are asymptotically negligible. Denote the remainders by

$$\tilde{\eta}_{\nu_n}^n = \sum_{i=2k_n}^{p_n+q_n} \left(\widehat{\varepsilon}_i^n - r_j^n\right) + \sum_{i=\nu_n(p_n+q_n)}^{N_t^n-k_n-j} \left(\widehat{\varepsilon}_i^n - r_j^n\right).$$

Let

$$S_n = \frac{1}{\sigma_n} \sum_{i=2k_n}^{N_t^n-k_n-j} \left(\widehat{\varepsilon}_i^n - r_j^n\right); \quad \tilde{S}_n = \frac{1}{\sigma_n} \sum_{k=1}^{\nu_n} \tilde{\eta}_k^n; \quad \bar{S}_n = \frac{1}{\sigma_n} \sum_{k=1}^{\nu_n-1} \bar{\eta}_k^n. \tag{86}$$

where

$$\sigma_n := \sqrt{\mathbf{Var}\left(\sum_{i=2k_n}^{N_t^n-k_n-j+1} \widehat{\varepsilon}_i^n\right)}.$$

From (71), we know $\sigma_n \asymp \Delta_n^{-1/2}$. Note that $S_n = \tilde{S}_n + \bar{S}_n$.

(a) Now we prove

$$\tilde{S}_n \xrightarrow{\mathbb{P}} 0. \tag{87}$$

Lemma D.3 implies

$$\mathbb{E}\left(\left(\tilde{\eta}_{\nu_n}^n/\sigma_n\right)^2\right) \le C\Delta_n(p_n+q_n) \to 0.$$

Similarly, we have for any $1 \le k \le \nu_n - 1$:

$$\mathbb{E}\left(\left(\tilde{\eta}_k^n\right)^2\right) \le Cp_n; \quad \mathbb{E}\left(\left(\tilde{\eta}_k^n\right)^2\right) \le Cq_n. \tag{88}$$

Let $\tilde{S}_n' = \frac{1}{\sigma_n}\sum_{k=1}^{\nu_n-1} \tilde{\eta}_k^n$. We prove $\tilde{S}_n' \xrightarrow{\mathbb{P}} 0$ given (81) and (82). The following is easy to get in view of (88) and $v > 2$:

$$\mathbb{E}\left(\left(\sum_{k=1}^{\nu_n-1} \tilde{\eta}_k^n\right)^2\right) = \left(\sum_{k=1}^{\nu_n-1} \mathbb{E}\left(\tilde{\eta}_k^n\right)^2\right) + 2\sum_{k=1}^{\nu_n-2}\sum_{k'>k}^{\nu_n-1} \mathbb{E}\left(\tilde{\eta}_k^n \tilde{\eta}_{k'}^n\right) \le C\nu_n q_n,$$

and it leads to

$$\mathbb{E}\left(\left(\tilde{S}_n'\right)^2\right) \le C\nu_n\Delta_n q_n \asymp q_n\left(p_n+q_n\right)^{-1} \xrightarrow{(81)} 0. \tag{89}$$

Therefore we have

$$\mathbb{E}\Big(\big(\tilde{S}_n\big)^2\Big) \le 2\left(\mathbb{E}\Big(\big(\tilde{S}'_n\big)^2\Big) + \mathbb{E}\Big(\big(\tilde{\eta}^n_{\nu_n}/\sigma_n\big)^2\Big)\right) \to 0. \tag{90}$$

This finishes the proof of (87).

(b) Now we prove the following Lindeberg condition :

$$\nu_n \mathbb{E}\Big(\mathbf{1}_{\{\bar{\eta}^n_k/\sigma_n > \epsilon\}}\big(\bar{\eta}^n_k/\sigma_n\big)^2\Big) \to 0, \text{ for any } \epsilon > 0. \tag{91}$$

We first have

$$\begin{aligned}
\mathbb{E}\Big(\big(\bar{\eta}^n_k\big)^4\Big) =& p_n \mathbb{E}\Big(\big(\widehat{\varepsilon}^n_i - r^n_j\big)^4\Big) + \sum_{i\neq i'} \mathbb{E}\Big(\big(\widehat{\varepsilon}^n_i - r^n_j\big)^2\big(\widehat{\varepsilon}^n_{i'} - r^n_j\big)^2\Big) \\
&+ \sum_{i\neq i'\neq i''} \mathbb{E}\Big(\big(\widehat{\varepsilon}^n_i - r^n_j\big)^2\big(\widehat{\varepsilon}^n_{i'} - r^n_j\big)\big(\widehat{\varepsilon}^n_{i''} - r^n_j\big)\Big) \\
&+ \sum_{i\neq i'\neq i''\neq i'''} \mathbb{E}\Big(\big(\widehat{\varepsilon}^n_i - r^n_j\big)\big(\widehat{\varepsilon}^n_{i'} - r^n_j\big)\big(\widehat{\varepsilon}^n_{i''} - r^n_j\big)\big(\widehat{\varepsilon}^n_{i'''} - r^n_j\big)\Big) \\
\le& C p_n^3 k_n.
\end{aligned}$$

To see the last inequality, we evaluate the following

$$\begin{aligned}
&\sum_{i>i'}\sum_{i'>i''}\sum_{i''>i'''}\sum_{i'''} \mathbb{E}\Big(\big(\widehat{\varepsilon}^n_i - r^n_j\big)\big(\widehat{\varepsilon}^n_{i'} - r^n_j\big)\big(\widehat{\varepsilon}^n_{i''} - r^n_j\big)\big(\widehat{\varepsilon}^n_{i'''} - r^n_j\big)\Big) \\
\le& \sum_{i>i'}\sum_{i'>i''}\sum_{i''>i'''}\left(Ck_n + \sum_{i'''<i''-2k_n-j} \mathbb{E}\Big(\big(\widehat{\varepsilon}^n_i - r^n_j\big)\big(\widehat{\varepsilon}^n_{i'} - r^n_j\big)\big(\widehat{\varepsilon}^n_{i''} - r^n_j\big)\big(\widehat{\varepsilon}^n_{i'''} - r^n_j\big)\Big)\right) \\
\le& C p_n^3 k_n.
\end{aligned}$$

Now we get

$$\begin{aligned}
\nu_n \mathbb{E}\Big(\mathbf{1}_{\{\bar{\eta}^n_k/\sigma_n > \epsilon\}}\big(\bar{\eta}^n_k/\sigma_n\big)^2\Big) &\le \nu_n \mathbb{E}\left(\mathbf{1}_{\{\bar{\eta}^n_k/\sigma_n > \epsilon\}} \frac{\big(\bar{\eta}^n_k/\sigma_n\big)^4}{\epsilon^2}\right) \\
&\le \frac{\nu_n \mathbb{E}\Big(\big(\bar{\eta}^n_k\big)^4\Big)}{\sigma_n^4 \epsilon^2} \le \frac{C\nu_n k_n p_n^3}{\sigma_n^4} \asymp k_n p_n^2 \Delta_n \overset{(83)}{\to} 0.
\end{aligned}$$

(c) Next, we show

$$\mathbb{E}\big(S_n^2\big) - \mathbb{E}\big(\bar{S}_n^2\big) \to 0. \tag{92}$$

To get (92), it suffices to show $\mathbb{E}\big(\bar{S}_n^2\big) < \infty$ since we have $\mathbb{E}\big(\tilde{S}_n^2\big) \to 0$ and $\mathbb{E}\big(S_n^2\big) - \mathbb{E}\big(\bar{S}_n^2\big) = \mathbb{E}\big(\tilde{S}_n^2\big) + 2\mathbb{E}\big(\tilde{S}_n \bar{S}_n\big)$. Apply the estimate in (88), we get $\mathbb{E}\big(\bar{S}_n^2\big) \le C\nu_n p_n \Delta_n$, which is bounded.

(d) Now we show

$$\left|\mathbb{E}(\exp(iu\bar{S}_n)) - \prod_{k=1}^{\nu_n - 1} \mathbb{E}\Big(\exp\big(iu\bar{\eta}^n_k/\sigma_n\big)\Big)\right| \le \frac{C\nu_n}{q_n^{v/2}} \overset{(84)}{\to} 0. \tag{93}$$

Let $\bar{\eta}^n_{\nu_n} = \bar{\eta}^n_0 = 0$. For $k = 0, \dots, \nu_n - 2$, let

$$\chi^n_k := \prod_{k'=0}^{k} \mathbb{E}\Big(\exp\left(iu\bar{\eta}^n_{k'}/\sigma_n\right)\Big)\mathbb{E}\left(\exp\left(iu\sum_{k''=k+1}^{\nu_n}\bar{\eta}^n_{k''}/\sigma_n\right)\right)$$
$$- \prod_{k'=0}^{k+1} \mathbb{E}\Big(\exp\left(iu\bar{\eta}^n_{k'}/\sigma_n\right)\Big)\mathbb{E}\left(\exp\left(iu\sum_{k''=k+2}^{\nu_n}\bar{\eta}^n_{k''}/\sigma_n\right)\right).$$

Rewrite

$$\mathbb{E}(\exp(iu\bar{S}_n)) - \prod_{k=1}^{\nu_n-1} \mathbb{E}\Big(\exp\left(iu\bar{\eta}^n_k/\sigma_n\right)\Big) = \sum_{k=0}^{\nu_n-2}\chi^n_k,$$

then apply Lemma A.1 yields $\left|\chi^n_k\right| \le C/q_n^{v/2}\ \forall k$ whence the inequality in (93). Therefore we see that the sequence $(\bar{\eta}^n_k)_{1\le k\le \nu_n-1}$ behave asymptomatically as if they are independent.

Now the proof of (78) is complete given (71), (87), (91), (92) and (93). If $\gamma > 1/v$, (79) follows from (69) and (78).

□

Lemma D.4. *Under the assumption of Theorem 3.3, we have*

$$\sqrt{N^n_t}\left(\mathrm{ReMeDI}(Y; j)^{\mathrm{HF}}_n - \mathrm{ReMeDI}(\varepsilon; j)^{\mathrm{HF}}_n\right) \xrightarrow{\mathbb{P}} 0. \tag{94}$$

Proof. By Cauchy-Schwarz inequality and the estimates in (53), we have

$$\mathbb{E}\left(\sqrt{N^n_t}\frac{\sum_{i=2k_n}^{N^n_t-k_n-j}\left|\Delta^{n,2k_n}_{i-2k_n}X\Delta^{n,k_n}_{i+j}X\right|}{N^n_t - 3k_n - j + 1}\right) \le Ck_n\sqrt{\Delta_n}.$$

Thus it suffices to prove

$$\sqrt{N^n_t}\frac{\sum_{i=2k_n}^{N^n_t-k_n-j}\left(\Delta^{n,2k_n}_{i-2k_n}X\Delta^{n,k_n}_{i+j}\varepsilon + \Delta^{n,2k_n}_{i-2k_n}\varepsilon\Delta^{n,k_n}_{i+j}X\right)}{N^n_t - 3k_n - j + 1} \xrightarrow{\mathbb{P}} 0,$$

since

$$\sqrt{N^n_t}\left(\mathrm{ReMeDI}(Y; j)^{\mathrm{HF}}_n - \mathrm{ReMeDI}(\varepsilon; j)^{\mathrm{HF}}_n\right)$$
$$= \sqrt{N^n_t}\frac{\sum_{i=2k_n}^{N^n_t-k_n-j}\left(\Delta^{n,2k_n}_{i-2k_n}X\Delta^{n,k_n}_{i+j}X + \Delta^{n,2k_n}_{i-2k_n}X\Delta^{n,k_n}_{i+j}\varepsilon + \Delta^{n,2k_n}_{i-2k_n}\varepsilon\Delta^{n,k_n}_{i+j}X\right)}{N^n_t - 3k_n - j + 1}.$$

We note by the independence of X, ε

$$\begin{aligned}
&\left|\sum_{i=2k_n}^{N_t^n-k_n-j}\sum_{k=2k_n}^{N_t^n-k_n-j}\mathbb{E}\left(\Delta_{i-2k_n}^{n,2k_n}X\Delta_{i+j}^{n,k_n}\varepsilon\Delta_{k-2k_n}^{n,2k_n}X\Delta_{k+j}^{n,k_n}\varepsilon\right)\right| \\
&\leq \sum_{i=2k_n}^{N_t^n-k_n-j}\sum_{k=2k_n}^{N_t^n-k_n-j}\left|\mathbb{E}\left(\Delta_{i-2k_n}^{n,2k_n}X\Delta_{k-2k_n}^{n,2k_n}X\right)\right|\left|\mathbb{E}\left(\Delta_{i+j}^{n,k_n}\varepsilon\Delta_{k+j}^{n,k_n}\varepsilon\right)\right| \\
&\leq Ck_n\Delta_n\sum_{i=2k_n}^{N_t^n-k_n-j}\sum_{k=2k_n}^{N_t^n-k_n-j}\left|\mathbb{E}\left(\Delta_{i+j}^{n,k_n}\varepsilon\Delta_{k+j}^{n,k_n}\varepsilon\right)\right| \\
&= Ck_n\Delta_n\sum_{i=2k_n}^{N_t^n-k_n-j}\left(\sum_{\{k:|i-k|\leq j+3k_n\}}\left|\mathbb{E}\left(\Delta_{i+j}^{n,k_n}\varepsilon\Delta_{k+j}^{n,k_n}\varepsilon\right)\right|+\sum_{\{k:|i-k|>j+3k_n\}}\left|\mathbb{E}\left(\Delta_{i+j}^{n,k_n}\varepsilon\Delta_{k+j}^{n,k_n}\varepsilon\right)\right|\right) \\
&\leq Ck_n^2.
\end{aligned}$$

The last inequality follows an application of Lemma A.1 and $v > 2$. This proves that

$$\mathbb{E}\left(\left(\sqrt{N_t^n}\frac{\sum_{i=2k_n}^{N_t^n-k_n-j}\Delta_{i-2k_n}^{n,2k_n}X\Delta_{i+j}^{n,k_n}\varepsilon}{N_t^n-3k_n-j+1}\right)^2\right)\leq Ck_n^2\Delta_n,$$

hence

$$\sqrt{N_t^n}\frac{\sum_{i=2k_n}^{N_t^n-k_n-j}\Delta_{i-2k_n}^{n,2k_n}X\Delta_{i+j}^{n,k_n}\varepsilon}{N_t^n-3k_n-j+1}\xrightarrow{\mathbb{P}}0.$$

Similarly we show

$$\sqrt{N_t^n}\frac{\sum_{i=2k_n}^{N_t^n-k_n-j}X_{i-2k_n}^{n,k_n}\varepsilon\Delta_{i+j}^{n,k_n}\varepsilon}{N_t^n-3k_n-j+1}\xrightarrow{\mathbb{P}}0.$$

This finish the proof of (94). □

Proof of Theorem 3.3. The proof simply follows from Theorem D.1 and Lemma D.4. □

E Proof of Theorem 3.4

We first introduce several notations. For each $2k_n \leq i \leq N_t^n - k_n - i_n - j$, denote

$$\begin{aligned}
&\tilde{\varepsilon}_i^n := \widehat{\varepsilon}_i^n - \mathrm{ReMeDI}(Y,j)_n^{\mathrm{HF}}; \quad \tilde{\varepsilon}_i^n(1) := \widehat{\varepsilon}_i^n - r_j^n; \quad \tilde{\varepsilon}_i^n(2) := \mathrm{ReMeDI}(Y,j)_n^{\mathrm{HF}} - r_j^n; \\
&\widehat{\Sigma}(\varepsilon;\iota_1,\iota_2)_j^n := \frac{1}{N_t^n-3k_n-j-i_n+1}\sum_{i=2k_n}^{N_t^n-k_n-i_n-j}\left(\tilde{\varepsilon}_i^n(\iota_1)\tilde{\varepsilon}_i^n(\iota_2)+2\sum_{k=1}^{i_n}\tilde{\varepsilon}_i^n(\iota_1)\tilde{\varepsilon}_{i+k}^n(\iota_2)\right), \iota_1,\iota_2\in\{1,2\}; \\
&\widehat{\Sigma}(\varepsilon)_j^n := \frac{1}{N_t^n-3k_n-j-i_n+1}\sum_{i=2k_n}^{N_t^n-k_n-i_n-j}\left(\left(\tilde{\varepsilon}_i^n\right)^2+2\sum_{k=1}^{i_n}\tilde{\varepsilon}_i^n\tilde{\varepsilon}_{i+k}^n\right).
\end{aligned}$$

Lemma E.1. *Assume all conditions of Theorem 3.4 hold. Then*

$$\widehat{\Sigma}(\varepsilon;1,2)_j^n + \widehat{\Sigma}(\varepsilon;2,1)_j^n + \widehat{\Sigma}(\varepsilon;2,2)_j^n \xrightarrow{\mathbb{P}} 0. \tag{95}$$

Proof. We first show

$$\mathbb{E}\left(\left(\mathrm{ReMeDI}(Y,j)_n^{\mathrm{HF}} - \mathrm{ReMeDI}(\varepsilon,j)_n^{\mathrm{HF}}\right)^2\right) \le Ck_n\Delta_n. \tag{96}$$

(96) can be obtained by replicating the arguments in the proof of Theorem 3.1. For the sake of completeness, we detail the proof. Let

$$\delta_i^n(1) = X_i^n - X_{i-2k_n}^n; \quad \delta_i^n(2) = X_{i+j}^n - X_{i+j+k_n}^n; \quad \tilde{\delta}_i^n(1) = \varepsilon_i^n - \varepsilon_{i-2k_n}^n; \quad \tilde{\delta}_i^n(2) = \varepsilon_{i+j}^n - \varepsilon_{i+j+k_n}^n.$$

Then we have

$$\begin{aligned}
\widehat{Y}_i^n &= \left(\delta_i^n(1) + \tilde{\delta}_i^n(1)\right)\left(\delta_i^n(2) + \tilde{\delta}_i^n(2)\right), \quad \widehat{\varepsilon}_i^n = \tilde{\delta}_i^n(1)\tilde{\delta}_i^n(2),\\
\widehat{Y}_i^n - \widehat{\varepsilon}_i^n &= \delta_i^n(1)\delta_i^n(2) + \tilde{\delta}_i^n(1)\delta_i^n(2) + \delta_i^n(1)\tilde{\delta}_i^n(2).
\end{aligned}$$

Apply the estimate (53), the fact ε has bounded moments and Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left(\left(\widehat{Y}_i^n - \widehat{\varepsilon}_i^n\right)^2\right) \le Ck_n\Delta_n. \tag{97}$$

Now (96) follows. Next, we show

$$\mathbb{E}\left(\left(\mathrm{ReMeDI}(\varepsilon,j)_n^{\mathrm{HF}} - r_j^n\right)^2\right) \le C\max\{\Delta_n, k_n^{-v/2}\}. \tag{98}$$

Apply (68), we have

$$\begin{aligned}
\sum_{i=2k_n}^{N_t^n-k_n-j}\sum_{k=0}^{N_t^n-k_n-j-i}\mathbb{E}\left(\left(\widehat{\varepsilon}_i^n - r_j^n\right)\left(\widehat{\varepsilon}_{i+k}^n - r_j^n\right)\right) &\le \sum_{i=2k_n}^{N_t^n-k_n-j}\sum_{k=0}^{N_t^n-k_n-j-i}\left(\mathbb{E}\left((\varepsilon_0\varepsilon_j - r_j)(\varepsilon_k\varepsilon_{k+j} - r_j)\right) + 3r_k^2 + \frac{C}{k_n^{v/2}}\right)\\
&\le C\left(\Delta_n^{-1} + \Delta_n^{-2}k_n^{-v/2}\right).
\end{aligned}$$

This proves (98). (96) and (98) imply $\mathbb{E}\left(\left(\tilde{\varepsilon}_i^n(2)\right)^2\right) \le \max\{k_n\Delta_n, k_n^{-v/2}\}$. Now apply the Cauchy-Schwarz inequality, we get

$$\mathbb{E}\left(\left|\widehat{\Sigma}(\varepsilon;1,2)_j^n\right| + \left|\widehat{\Sigma}(\varepsilon;2,1)_j^n\right| + \left|\widehat{\Sigma}(\varepsilon;2,2)_j^n\right|\right) \le Ci_n\max\{\sqrt{k_n\Delta_n}, k_n^{-v/4}\} \to 0.$$

□

Lemma E.2. *Under the assumptions of Theorem 3.4, we have*

$$\widehat{\Sigma}(\varepsilon;1,1)_j^n \xrightarrow{\mathbb{P}} \Sigma_j. \tag{99}$$

Proof. We first show

$$\mathbb{E}\left(N_t^n\left(\mathrm{ReMeDI}(\varepsilon;j)_n^{\mathrm{HF}}-r_j^n\right)^2\right)-\widehat{\Sigma}(\varepsilon;1,1)_j^n\to 0. \tag{100}$$

Let $\bar{\varepsilon}_{i,i'}^n=\left(\widehat{\varepsilon}_i^n-r_j^n\right)\left(\widehat{\varepsilon}_{i'}^n-r_j^n\right)-\mathbb{E}\left(\left(\widehat{\varepsilon}_i^n-r_j^n\right)\left(\widehat{\varepsilon}_{i'}^n-r_j^n\right)\right)$. First, we apply Lemma A.1 and get an estimate $\mathbb{E}\left(\left(\sum_{i=2k_n}^{N_t^n-k_n-i_n-j}\bar{\varepsilon}_{i,i}^n\right)^2\right)\le C\Delta_n^{-1}k_n$, which implies

$$\frac{1}{N_t^n-3k_n-i_n-j+1}\sum_{i=2k_n}^{N_t^n-k_n-i_n-j}\bar{\varepsilon}_{i,i}^n\xrightarrow{\mathbb{P}}0. \tag{101}$$

Similarly, we can prove

$$\frac{1}{N_t^n-3k_n-i_n-j+1}\sum_{i=2k_n}^{N_t^n-k_n-i_n-j}\sum_{k=1}^{i_n}\left(\bar{\varepsilon}_{i,i+k}^n+\bar{\varepsilon}_{i+k,i}^n\right)\xrightarrow{\mathbb{P}}0. \tag{102}$$

Another application of Lemma A.1 yields

$$\frac{1}{N_t^n-3k_n-i_n-j+1}\mathbb{E}\left(\sum_{i=2k_n}^{N_t^n-k_n-i_n-j}\sum_{k=i_n+1}^{N_t^n-k_n-i_n-j-i}\left(\widehat{\varepsilon}_i^n-r_j^n\right)\left(\widehat{\varepsilon}_{i+k}^n-r_j^n\right)\right)\le Ci_n^{1-\frac{v}{2}}. \tag{103}$$

(101), (102) and (103) imply

$$\widehat{\Sigma}(\varepsilon;1,1)_j^n-\frac{1}{N_t^n-3k_n-i_n-j+1}\sum_{i=2k_n}^{N_t^n-k_n-i_n-j}\sum_{i'=2k_n}^{N_t^n-k_n-i_n-j}\mathbb{E}\left(\left(\widehat{\varepsilon}_i^n-r_j^n\right)\left(\widehat{\varepsilon}_{i'}^n-r_j^n\right)\right)\xrightarrow{\mathbb{P}}0. \tag{104}$$

However, the remainders are negligible due to the conditions on i_n:

$$\mathbb{E}\left(N_t^n\left(\mathrm{ReMeDI}(\varepsilon;j)_n^{\mathrm{HF}}-r_j^n\right)^2\right)-\frac{\sum_{i=2k_n}^{N_t^n-k_n-i_n-j}\sum_{i'=2k_n}^{N_t^n-k_n-i_n-j}\mathbb{E}\left(\left(\widehat{\varepsilon}_i^n-r_j^n\right)\left(\widehat{\varepsilon}_{i'}^n-r_j^n\right)\right)}{N_t^n-3k_n-i_n-j+1}\to 0. \tag{105}$$

Now (100) follows from (104) and (105). Note that (68) implies

$$\mathbb{E}\left(N_t^n\left(\mathrm{ReMeDI}(\varepsilon;j)_n^{\mathrm{HF}}-r_j^n\right)^2\right)\to\Sigma_j. \tag{106}$$

Now the result follows from (100) and (106). □

Proof of Theorem 3.4. The following is immediate to obtain

$$\bar{\varepsilon}_i^n=\bar{\varepsilon}_i^n(1)+\bar{\varepsilon}_i^n(2);\quad \widehat{\Sigma}(\varepsilon)_j^n=\sum_{\iota_1=1}^{2}\sum_{\iota_2=1}^{2}\widehat{\Sigma}(\varepsilon;\iota_1,\iota_2)_j^n.$$

Then by Lemma E.1, Lemma E.2 and Theorem 3.3, it suffices to show

$$\widehat{\Sigma}(Y)^n_j - \widehat{\Sigma}(\varepsilon)^n_j \xrightarrow{\mathbb{P}} 0. \tag{107}$$

Now we will show

$$\mathbb{E}\left(\left|\widehat{\Sigma}(Y)^n_j - \widehat{\Sigma}(\varepsilon)^n_j\right|\right) \leq C\left(i_n(k_n\Delta_n)^{1/2}\right). \tag{108}$$

Once this is proved, (107) follow since $\sqrt{k_n\Delta_n} = o_p\left(\Delta_n^{2/5}\right), i_n \asymp \Delta_n^{-1/5}$. For any $k \geq 0$,

$$\widetilde{Y}^n_i\widetilde{Y}^n_{i+k} - \tilde{\varepsilon}^n_i\tilde{\varepsilon}^n_{i+k} = \left(\widehat{Y}^n_{i+k} - \widehat{\varepsilon}^n_{i+k}\right)\left(\widehat{\varepsilon}^n_i - \text{ReMeDI}(Y;j)^{\text{HF}}_n\right) + \left(\widehat{Y}^n_i - \widehat{\varepsilon}^n_i\right)\left(\widehat{Y}^n_{i+k} - \text{ReMeDI}(Y;j)^{\text{HF}}_n\right).$$

Now (97), Cauchy-Schwarz inequality and the (trivial) facts that $\mathbb{E}\left(\left(\widehat{\varepsilon}^n_i - \text{ReMeDI}(Y;j)^{\text{HF}}_n\right)^2\right) \leq C$, $\mathbb{E}\left(\left(\widehat{Y}^n_{i+k} - \text{ReMeDI}(Y;j)^{\text{HF}}_n\right)^2\right) \leq C$ imply

$$\mathbb{E}\left(\left|\widetilde{Y}^n_i\widetilde{Y}^n_{i+k} - \tilde{\varepsilon}^n_i\tilde{\varepsilon}^n_{i+k}\right|\right) \leq C\sqrt{k_n\Delta_n}.$$

Now (108) is proved and this completes the proof.

□

F Proof of the results in Section 5

Proof of Theorem 5.2. This follows from Theorem 3.3, Theorem 3.4 and the delta method. □

Proof of Corollary 5.1. (69), (96) and (98) yield for any j

$$\left|\text{ReMeDI}(Y;j)^{\text{HF}}_n - r_j\right| = O_p\left(\max\{(k_n\Delta_n)^{1/2}, k_n^{-v/2}\}\right).$$

The consistency result (41) follows from (40). □

Proof of Theorem 5.1. Replicate the proof of Theorem 3.1, we obtain a similar result of (62):

$$\mathbb{E}\left(n^{-2}\left(\sum_{i=2k_n}^{n-k_n-j}\left(\widehat{Y}(j)_i - r_j\right)\right)^2\right) \leq \max\{k_n^{-v}, k_n/n\},$$

whence $\text{ReMeDI}(Y,j)^{\text{FF}}_n - r_j = O_p\left(\max\{k_n^{-v/2}, \sqrt{k_n/n}\}\right)$. Now the results follows from the asymptotic conditions on ℓ^{HF}_n. □

G Autocorrelation patterns of bid-ask spread of INTC and KO, daily estimation

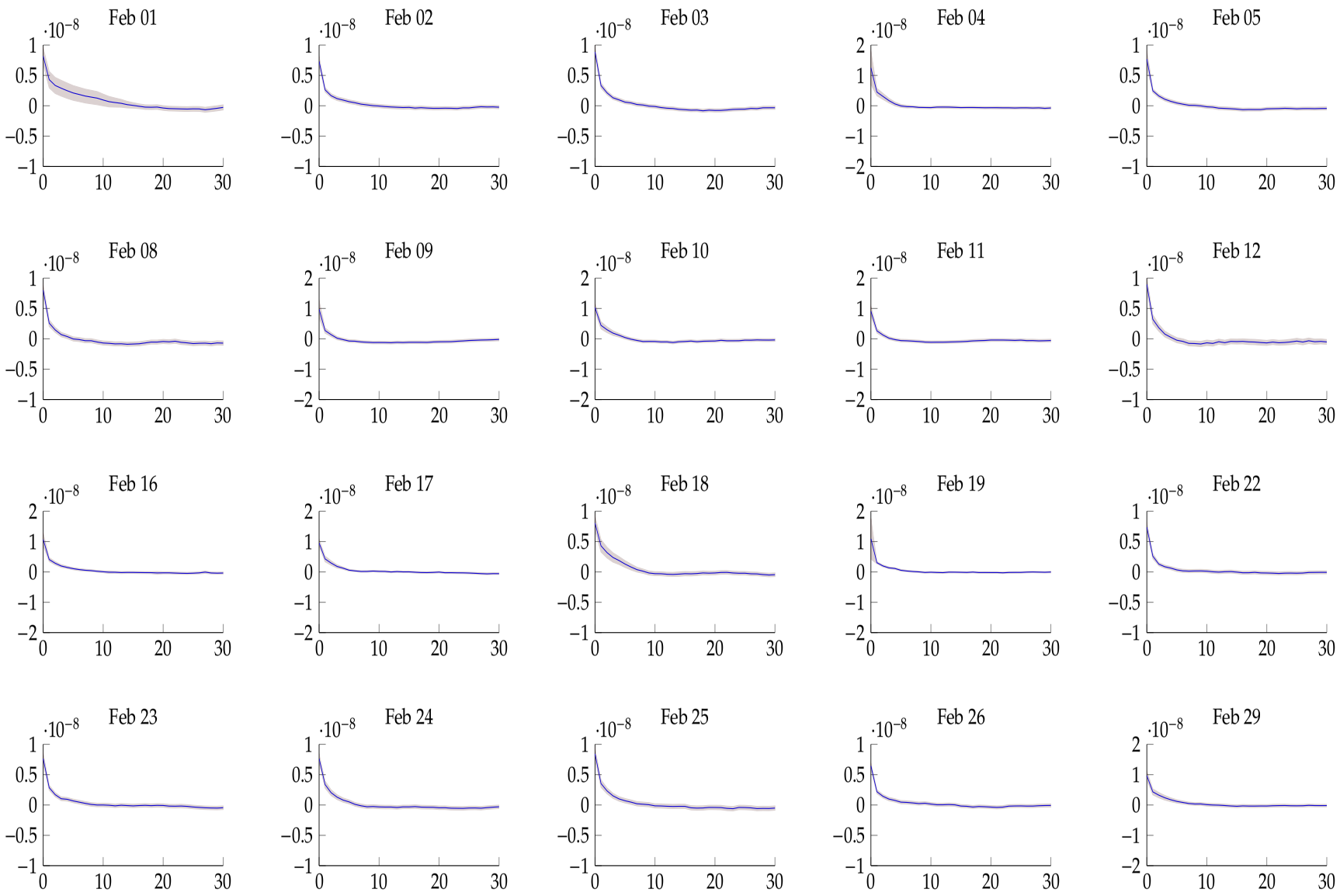

Figure 18: Autocovariances of bid-ask spread for INTC. In each of the 20 trading days of February, 2016, transaction prices between 9:35 AM and 4:00 PM are collected. The ReMeDI estimates (with $k_n = 10$) of the autocovariances (up to 30 lags) are obtained. The shaded areas are the 95% confidence intervals. The tuning parameter to computer the confidence intervals $i_n = 6$..

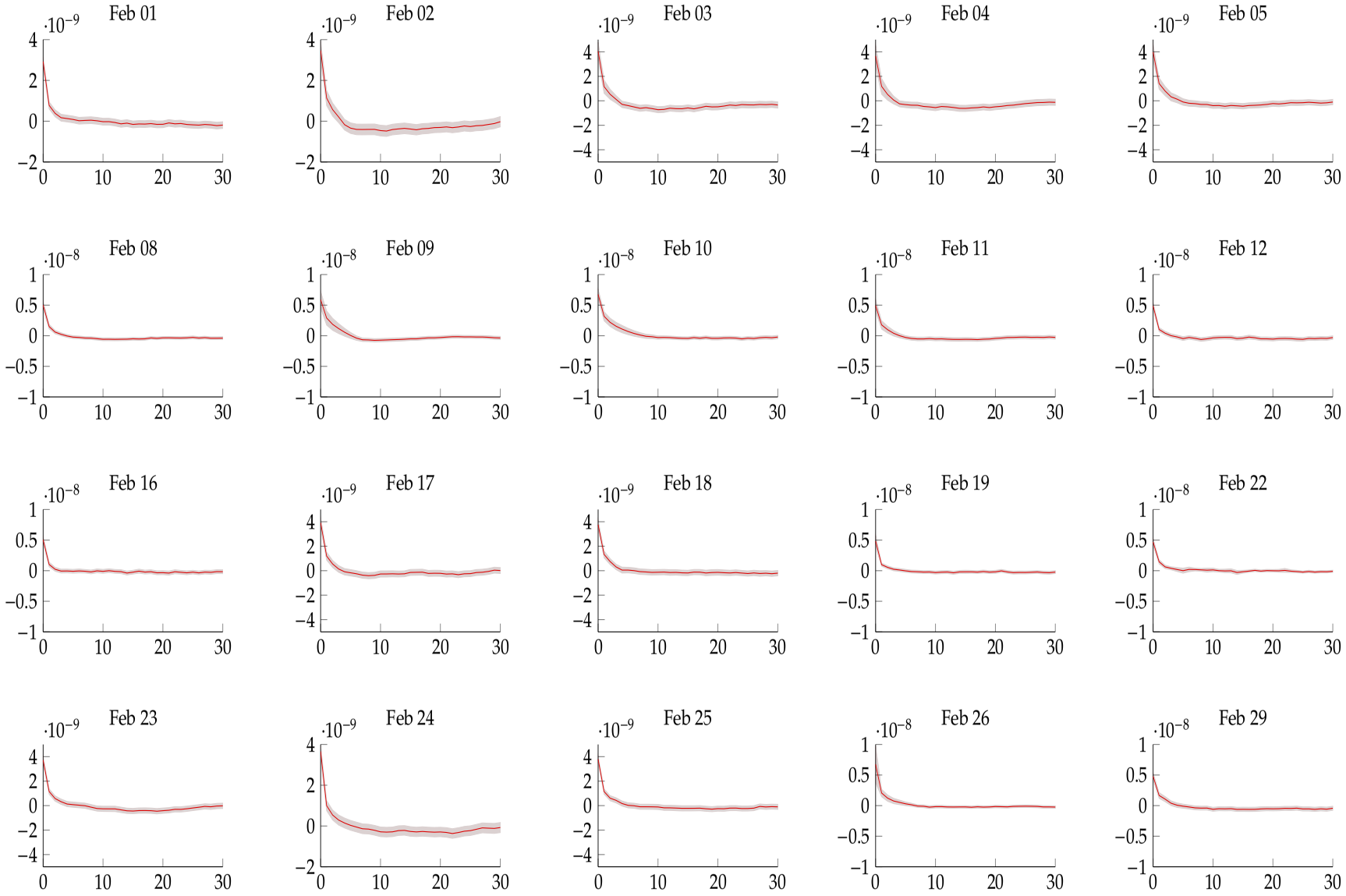

Figure 19: Autocovariances of bid-ask spread for KO. In each of the 20 trading days of February, 2016, transaction prices between 9:35 AM and 4:00 PM are collected. The ReMeDI estimates (with $k_n = 10$) of the autocovariances (up to 30 lags) are obtained. The shaded areas are the 95% confidence intervals. The tuning parameter to computer the confidence intervals $i_n = 6$.

References

Aït-Sahalia, Y. and J. Jacod (2014): *High-Frequency Financial Econometrics*, Princeton University Press.

Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2005): "How often to sample a continuous-time process in the presence of market microstructure noise," *Review of Financial Studies*, 18, 351–416.

——— (2011): "Ultra high frequency volatility estimation with dependent microstructure noise," *Journal of Econometrics*, 160, 160–175.

Ait-Sahalia, Y. and D. Xiu (2017): "A hausman test for the presence of market microstructure noise in high frequency data," *Journal of Econometrics, forthcoming*.

Bandi, F. M., D. Pirino, and R. Reno (2017): "EXcess Idle Time," *Econometrica, forthcoming*.

Bandi, F. M. and J. R. Russell (2006): "Separating microstructure noise from volatility," *Journal of Financial Economics*, 79, 655–692.

——— (2008): "Microstructure noise, realized variance, and optimal sampling," *Review of Economic Studies*, 75, 339–369.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2008): "Designing realized kernels to measure the ex post variation of equity prices in the presence of noise," *Econometrica*, 76, 1481–1536.

Barndorff-Nielsen, O. E. and N. Shephard (2004): "Power and bipower variation with stochastic volatility and jumps," *Journal of Financial Econometrics*, 2, 1–37.

——— (2006): "Econometrics of testing for jumps in financial economics using bipower variation," *Journal of Financial Econometrics*, 4, 1–30.

Bogousslavsky, V. (2016): "Infrequent rebalancing, return autocorrelation, and seasonality," *Journal of Finance*, 71, 2967–3006.

Bollerslev, T., J. Li, and Y. Xue (2018): "Volume, Volatility and Public News Announcements," *Review of Economic Studies, forthcoming*.

Bradley, R. C. (2005): "Basic properties of strong mixing conditions. A survey and some open questions," *Probability Surveys*, 2, 107–144.

Campbell, J. Y., S. J. Grossman, and J. Wang (1993): "Trading volume and serial correlation in stock returns," *Quarterly Journal of Economics*, 108, 905–939.

Chan, K. C., W. G. Christie, and P. H. Schultz (1995): "Market structure and the intraday pattern of bid-ask spreads for NASDAQ securities," *Journal of Business*, 35–60.

Chan, L. K. and J. Lakonishok (1995): "The behavior of stock prices around institutional trades," *Journal of Finance*, 50, 1147–1174.

Chen, X., O. Linton, S. Schneeberger, and Y. Yi (2017a): "Semiparametric Estimation of the Bid-Ask Spread in Extended Roll Models," .

Chen, X., O. Linton, and Y. Yi (2017b): "Semiparametric identification of the bid–ask spread in extended Roll models," *Journal of Econometrics*, 200, 312–325.

Choi, J. Y., D. Salandro, and K. Shastri (1988): "On the estimation of bid-ask spreads: Theory and evidence," *Journal of Financial and Quantitative Analysis*, 23, 219–230.

Da, R. and D. Xiu (2017): "When Moving-Average Models Meet High-Frequency Data: Uniform Inference on Volatility," Tech. rep.

Diebold, F. X. and G. Strasser (2013): "On the correlation structure of microstructure noise: A financial economic approach," *Review of Economic Studies*, 80, 1304–1337.

Ellis, K., R. Michaely, and M. O'Hara (2000): "The accuracy of trade classification rules: Evidence from Nasdaq," *Journal of Financial and Quantitative Analysis*, 35, 529–551.

Grossman, S. J. and M. H. Miller (1988): "Liquidity and market structure," *Journal of Finance*, 43, 617–633.

Hall, P. and C. C. Heyde (1980): *Martingale Limit Theory and Its Application*, Academic Press.

Hansen, P. R. and A. Lunde (2006): "Realized variance and market microstructure noise," *Journal of Business & Economic Statistics*, 24, 127–161.

Harris, L. (1990): "Estimation of stock price variances and serial covariances from discrete observations," *Journal of Financial and Quantitative Analysis*, 25, 291–306.

Hasbrouck, J. (1993): "Assessing the quality of a security market: A new approach to transaction-cost measurement," *Review of Financial Studies*, 6, 191–212.

——— (2004): "Liquidity in the futures pits: Inferring market dynamics from incomplete data," *Journal of Financial and Quantitative Analysis*, 39, 305–326.

——— (2007): *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*, Oxford University Press.

——— (2009): "Trading costs and returns for US equities: Estimating effective costs from daily data," *Journal of Finance*, 64, 1445–1477.

Hasbrouck, J. and T. S. Ho (1987): "Order Arrival, Quote Behavior, and the Return-Generating Process," *Journal of Finance*, 42, 1035–1048.

Hasbrouck, J. and G. Sofianos (1993): "The trades of market makers: An empirical analysis of NYSE specialists," *Journal of Finance*, 48, 1565–1593.

Hautsch, N. and M. Podolskij (2013): "Preaveraging-Based Estimation of Quadratic Variation in the Presence of Noise and Jumps: Theory, Implementation, and Empirical Evidence," *Journal of Business & Economic Statistics*, 31, 165–183.

Hendershott, T., C. Jones, and A. J. Menkveld (2013): "Implementation shortfall with transitory price effects," *High Frequency Trading; New Realities for Trades, Markets and Regulators, Easley, D., M. Lopez de Prado, and M. O?Hara (editors), Risk Books (London: 2013)*.

Hendershott, T. and A. J. Menkveld (2014): "Price pressures," *Journal of Financial Economics*, 114, 405–423.

Hendershott, T., R. Praz, A. J. Menkveld, and M. S. Seasholes (2018): "Asset price dynamics with limited attention," .

Ho, T. and H. R. Stoll (1981): "Optimal dealer pricing under transactions and return uncertainty," *Journal of Financial Economics*, 9, 47–73.

Huang, R. D. and H. R. Stoll (1997): "The components of the bid-ask spread: A general approach," *Review of Financial Studies*, 10, 995–1034.

Ibragimov, I. A. (1962): "Some limit theorems for stationary processes," *Theory of Probability & Its Applications*, 7, 349–382.

Jacod, J., Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter (2009): "Microstructure noise in the continuous case: the pre-averaging approach," *Stochastic Processes and their Applications*, 119, 2249–2276.

Jacod, J., Y. Li, and X. Zheng (2015): "Estimating the Integrated Volatility When Microstructure Noise is Dependent and Observation Times are Irregular," Tech. rep.

——— (2017): "Statistical properties of microstructure noise," *Econometrica*, 85, 1133 – 1174.

Jacod, J. and P. E. Protter (2011): *Discretization of Processes*, vol. 67, Springer Science & Business Media.

Jacod, J. and A. N. Shiryaev (2003): *Limit Theorems for Stochastic Processes*, vol. 288, Springer-Verlag Berlin.

Jain, P. C. and G.-H. Joh (1988): "The dependence between hourly prices and trading volume," *Journal of Financial and Quantitative Analysis*, 23, 269–283.

Lee, C. and M. J. Ready (1991): "Inferring trade direction from intraday data," *Journal of Finance*, 46, 733–746.

Li, Z. M., R. J. Laeven, and M. H. Vellekoop (2017): "Dependent microstructure noise and integrated volatility estimation from high-frequency data," https://arxiv.org/abs/1704.08964.

Li, Z. M. and O. Linton (2018): "Robust measures of microstructure noise," Tech. rep.

Lin, J.-C., G. C. Sanger, and G. G. Booth (1995): "Trade size and components of the bid-ask spread," *Review of Financial Studies*, 8, 1153–1183.

Madhavan, A., M. Richardson, and M. Roomans (1997): "Why do security prices change? A transaction-level analysis of NYSE stocks," *Review of Financial Studies*, 10, 1035–1064.

McInish, T. H. and R. A. Wood (1990): "An analysis of transactions data for the Toronto Stock Exchange: Return patterns and end-of-the-day effect," *Journal of Banking & Finance*, 14, 441–458.

——— (1992): "An analysis of intraday patterns in bid/ask spreads for NYSE stocks," *Journal of Finance*, 47, 753–764.

Mokkadem, A. (1988): "Mixing properties of ARMA processes," *Stochastic Processes and their Applications*, 29, 309–315.

Podolskij, M. and M. Vetter (2009): "Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps," *Bernoulli*, 15, 634–658.

Roll, R. (1984): "A simple implicit measure of the effective bid-ask spread in an efficient market," *Journal of Finance*, 39, 1127–1139.

Sadka, R. (2006): "Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk," *Journal of Financial Economics*, 80, 309–349.

Stoll, H. R. (1989): "Inferring the components of the bid-ask spread: theory and empirical tests," *Journal of Finance*, 44, 115–134.

——— (2000): "Friction," *Journal of Finance*, 55, 1479–1514.

——— (2003): "Market microstructure," *Handbook of the Economics of Finance*, 1, 553–604.

Todorov, V. and G. Tauchen (2011): "Volatility jumps," *Journal of Business & Economic Statistics*, 29, 356–371.

Wood, R. A., T. H. McInish, and J. K. Ord (1985): "An investigation of transactions data for NYSE stocks," *Journal of Finance*, 40, 723–739.

Xiu, D. (2010): "Quasi-maximum likelihood estimation of volatility with high frequency data," *Journal of Econometrics*, 159, 235–250.

Zhang, L. (2006): "Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach," *Bernoulli*, 12, 1019–1043.

Zhang, L., P. A. Mykland, and Y. Aït-Sahalia (2005): "A tale of two time scales: determining integrated volatitility with noisy high-frequency data," *Journal of the American Statistical Association*, 100, 1394–1411.