UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Institute for New Economic Thinking

Cambridge Working Papers in Economics

Cambridge-INET Working Paper Series No: 2019/24

Cambridge Working Papers in Economics: 1908

A REMEDI FOR MICROSTRUCTURE NOISE

Z. Merrick Li
(University of Cambridge)

Oliver Linton
(University of Cambridge)

We introduce the Realized moMents of Disjoint Increments (ReMeDI) paradigm to measure microstructure noise (the deviation of the observed asset prices from the fundamental values caused by market imperfections). We propose consistent estimators of arbitrary moments of the microstructure noise process based on highfrequency data, where the noise process could be serially dependent, endogenous, and nonstationary. We characterize the limit distributions of the proposed estimators and construct confidence intervals under infill asymptotics. Our simulation and empirical studies show that the ReMeDI approach is very effective to measure the scale and the serial dependence of microstructure noise. Moreover, the estimators are quite robust to model specifications, sample sizes and data frequencies.

Keywords: Microstructure noise, semimartingale, serial dependence, stable convergence, mixing sequence, infill asymptotics, finite sample bias

A ReMeDI for Microstructure Noise *

Z. Merrick Li† Oliver Linton‡

September 8, 2020

Abstract

We introduce the Realized moMents of Disjoint Increments (ReMeDI) paradigm to measure microstructure noise (the deviation of the observed asset prices from the fundamental values caused by market imperfections). We propose consistent estimators of arbitrary moments of the microstructure noise process based on high-frequency data, where the noise process could be serially dependent, endogenous, and nonstationary. We characterize the limit distributions of the proposed estimators and construct confidence intervals under infill asymptotics. Our simulation and empirical studies show that the ReMeDI approach is very effective to measure the scale and the serial dependence of microstructure noise. Moreover, the estimators are quite robust to model specifications, sample sizes and data frequencies.

KEYWORDS: Microstructure noise, semimartingale, serial dependence, stable convergence, mixing sequence, infill asymptotics, finite sample bias

*Part of the paper was circulated under various titles, including, "A simple measure of microstructure noise"and "Robust measures of microstructure noise". We benefit from discussions with Yacine Aït-Sahalia, Torben Andersen, Peter Boswijk, Rob Engle, Christian Gouriéroux, Peter Reinhard Hansen, Jean Jacod, Ilze Kalnina, Frank Kleibergen, Eben Lazarus, Yingying Li, Albert Menkveld, Per Mykland, George Tauchen, Michel Vellekoop, Bas Werker, Dacheng Xiu, Xiye Yang, Xinghua Zheng, and seminar participants at the 2017 North American Summer Meeting (St. Louis, June 2017), 2017 European Meeting (Lisbon, August 2017) and 2018 China Meeting (Shanghai, June 2018) of the Econometric Society, International Conference on Quantitative Finance and Financial Econometrics (Marseille, June 2018), the 12th Annual SoFiE Conference (Shanghai, June 2019), the Econometric Society and Bocconi University Virtual World Congress (August, 2020), Hong Kong University of Science and Technology, NYU Shanghai, University of Amsterdam and University of Cambridge. The project is partially sponsored by the Keynes Fund (JHUL).

†Faculty of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom. Email: z.merrick.li@gmail.com

‡Faculty of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom. Email: obl20@cam.ac.uk.

1 Introduction

Economic time series are often modelled as the sum of a latent process obtained from an underlying economic model and another term that reflects a variety of adjustments to or departures from the frictionless theoretical model, thus

$$\underbrace{Y}_{\text{observed series}} = \underbrace{X}_{\text{underlying process}} + \underbrace{\varepsilon}_{\text{deviation}} . \tag{1}$$

The two processes X and ε are generated by different mechanisms, and can have quite distinct statistical properties and economic interpretations. Both quantities may be of interest as they give interpretation of some underlying economic theory and its relevance for the observed data. However, since only the sum process Y is observable, this makes estimation and inference challenging.

We are concerned with applications of this framework in financial markets where the observed asset price[1] (Y) subsumes both *market microstructure noise* (ε) and the *efficient price* (or *fundamental value)* (X). The fundamental theorem of asset pricing says that X should be a semimartingale process (Delbaen and Schachermayer (1994)). In practice however, many market frictions, such as: transaction costs, price discreteness, inventory holdings, information asymmetry, measurement errors, may cause observed prices to deviate from this ideal price. One may also want to allow for temporary mispricing (French and Roll (1986)) or fad effects (Lehmann (1990)); see also O'Hara (1995), Hasbrouck (2007) and Foucault et al. (2013) for insightful reviews. A lot of early work proceeded on the basis that the microstructure noise process was i.i.d., but recently this assumption has been shown to be too strong; both theoretically and empirically the microstructure noise may exhibit rich dynamics depending on its origin. If the microstructure effects are negligible, the observed price should be close to the efficient price and be unpredictable. Therefore, the dispersion and persistence of the microstructure noise serve as natural measures of market quality. Market quality is of concern to regulators and practitioners as well as academics; proxies for market quality are widely used in empirical analysis, see Linton and Mahmoodzadeh (2018). In the working paper Li and Linton (2019), we proposed two market liquidity measures, called IBAS and ABAS, that were defined in terms of the autocovariance function of the noise process. Such liquidity measures are robust to the pattern of order flows, and have an intuitive economic interpretation.

We introduce a general econometric approach to measuring microstructure noise in a nonparametric setting. Specifically, we propose a new estimator of the moments

[1] By *price* it always means the *logarithmic price* in this paper unless stated otherwise.

of a general dependent noise process based on the observed noisy high-frequency transaction prices; we call our estimator the *Realized moMents of Disjoint Increments* (ReMeDI). The estimation method is based on the differencing paradigm, which is widely used in microeconometrics to eliminate nuisance parameters, see, e.g., Athey and Imbens (2006). We build on the general setup introduced in the seminal work of Jacod et al. (2017). Specifically, we assume that the underlying efficient price follows a semimartingale, which may accommodate stochastic volatility, jumps, etc. We allow the microstructure noise to be weakly dependent and to have a serial correlation of an unknown form that may decay at an algebraic rate; this may capture, for instance, the effects of clustered (or hidden) order flows or herding (Park and Sabourian (2011)). The microstructure noise is allowed to have time-varying and stochastic heteroskedasticity, which allows for intraday variation in the scale of the noise. The general setting we consider allows for random and endogenous observation schemes. We develop estimators of arbitrary moments of the microstructure noise; this includes the autocovariance function of powers of the noise process as well as other quantities of interest. We derive the stable convergence in law of the estimated quantities as the sample size increases on a given domain. We provide a consistent estimator of the asymptotic variance that allows us to quantify the accuracy of our estimator.

We present some simulation studies comparing the ReMeDI approach with the method of Jacod et al. (2017). We find that the ReMeDI approach is relatively robust to: the data frequency, the sample size, the tuning parameter, and to model specification. We provide an empirical study on an individual stock price, which reveals that the microstructure noise has non-trivial serial dependence, but that the dependence structure falls short of being long memory. This is consistent with leading microstructure models,[2] and differs from the findings in Jacod et al. (2017).

The robustness of the ReMeDI approach as demonstrated in our simulation and empirical studies has an intuitive explanation. The differencing method works because the increments of X over disjoint intervals (the efficient returns) are small and/or uncorrelated, and what remains is attributed to ε. This property distinguishes the ReMeDI approach from alternative high-frequency estimators that rely structurally on the *infill asymptotics*.

1.1 Related literature

There are a number of methods for estimation of the moments of noise and the parameters of the efficient price. Specifically: the two-scale/multi-scale realized volatility

[2]For example, Hasbrouck and Ho (1987), Choi et al. (1988) and Huang and Stoll (1997) model the probability of order reversal, and microstructure noise becomes an AR(1) process.

by Zhang et al. (2005), Zhang (2006), Aït-Sahalia et al. (2011); the optimal-sampling realized variance by Bandi and Russell (2006, 2008); the maximum likelihood estimators by Aït-Sahalia et al. (2005), Xiu (2010); the pre-averaging method developed in Podolskij and Vetter (2009), Jacod et al. (2009); and the realized kernel by Hansen and Lunde (2006), Barndorff-Nielsen et al. (2008). Most of this literature only considers i.i.d. microstructure noise.

Several recent papers explore richer microstructure models by allowing for autocorrelated noise. The estimators of the second moments of noise in Da and Xiu (2019) and Li et al. (2020) are by-products of the integrated volatility estimators in the presence of autocorrelated noise. In a recent seminal paper, Jacod et al. (2017) introduced the first feasible procedure, called the *local averaging* (LA) method, to estimate arbitrary moments of microstructure noise using high-frequency data. They also introduced a general framework allowing for a stochastic observation scheme and a microstructure noise with a semimartingale "size process". We follow their general setup and derive asymptotic properties of our estimators under this general framework. We differentiate our paper from Jacod et al. (2017) as follows. First, the ReMeDI method is based on *differencing*; while the LA method is based on deviations from *local averages*, both ideas are widely used in other contexts such as panel data and semiparametric estimation to eliminate nuisance parameters, see Yatchew (1997) and Athey and Imbens (2006). Second, the ReMeDI approach works beyond the infill framework. Specifically, in the working paper version, Li and Linton (2019), we proved that the ReMeDI estimator is consistent and has an associated CLT in a long-span, non-infill setting. In this case, the method works provided the efficient price is a martingale in which case its increments are uncorrelated at any horizon. The LA method, however, is inconsistent when applied to low-frequency data. Next, the finite sample performance of the LA estimators heavily depends on the sample size and the noise-to-signal ratio (the ratio of noise variance to the integrated volatility of the efficient price), see an analysis in Jacod et al. (2017). This may cause many issues in the implementations with real data.[3] The bias of the ReMeDI estimators by contrast only depends on the slope of the microstructure autocovariance function, and in short memory contexts this bias can be very small. Last, the ReMeDI approach has another two advantages in real implementations: it is computationally very efficient,[4] and it is very robust to a wide range of tuning parameters.

[3]One can easily verify the following scenarios by simulation: (1) the LA estimator may report positive autocovariances when the true noise process is uncorrelated or even negatively autocorrelated; (2) the LA estimator has larger bias and variance if there are bursts of volatility in the efficient price process, e.g., when the volatility process jumps; (3) the LA estimator gives very different estimates over two samples where the noise processes are identical but the efficient prices have different variances.

[4]For example, the LA (ReMeDI) takes 99.77% (0.23%) of the CPU time to estimate the variance of the noise using noisy price data from a random walk plus AR(1) noise model, based on 1,000 simulated

Our paper introduces an econometric approach to richer microstructure models. It aims to integrate the market microstructure and financial econometrics literature. 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 implications for integrated volatility estimation. Li et al. (2016), Chaker (2017) and Clinet and Potiron (2017) model the microstructure noise as a parametric function of the observable trading information and develop efficient volatility estimators. Bandi et al. (2017) develop a novel measure of the staleness of stock returns under the infill asymptotic framework. 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 high-frequency econometric techniques to identify jumps. Da and Xiu (2019) advocate the quasi-maximum likelihood approach to estimate both the volatility and the autocovariances of moving-average microstructure noise.

2 Continuous-Time Framework and Assumptions

We follow the general framework of Jacod et al. (2017) to specify the continuous-time efficient price process, the observation scheme, and the microstructure noise. We have almost the same regularity conditions as Jacod et al. (2017).[5] Hence, for brevity of exposition, we put some details of the specifications in Appendix A.

2.1 Efficient price process

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

$$X_t = X_0 + \int_0^t b_s \mathrm{d}s + \int_0^t \sigma_s \mathrm{d}W_s + \left(\vartheta \mathbf{1}_{\{|\vartheta|\leq 1\}}\right) \star (\mathfrak{p} - \mathfrak{q})_t + \left(\vartheta \mathbf{1}_{\{|\vartheta|>1\}}\right) \star \mathfrak{p}_t, \quad (2)$$

where $W, \mathfrak{p}$ are a Wiener process and a Poisson random measure on $\mathbb{R}_+$and E respectively. Here, $(E, \mathcal{E})$ is a measurable Polish space on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ and the predictable compensator of $\mathfrak{p}$ is $\mathfrak{q}(\mathrm{d}s, \mathrm{d}z) = \mathrm{d}s \otimes \lambda(\mathrm{d}z)$ for some given σ-finite measure on $(E, \mathcal{E})$, see Jacod and Shiryaev (2003) for detailed introduction of the last two integrals. Further regularity conditions on X are discussed in Assumption (H) in Appendix A. Note that the setting of the efficient price is very general, and it allows for

samples of size 23,400.

[5]The only difference is that we have more restriction on the parameter that controls the degree of serial dependence of the stationary noise, see Remark 2.1.

stochastic volatility and jumps in both the price and volatility processes.

2.2 Observation scheme

We assume a triangular array structure. For each n, let $\{t_i^n : i \in \mathbb{N}_+\}$ be a sequence of random finite observed times (usually when a transaction occurs) with $0 = t_0^n < t_1^n < \ldots$, where $\mathbb{N}_+$ is the set of nonnegative integers. We denote

$$n_t := \sum_{i \geq 0} \mathbf{1}_{\{t_i^n \leq t\}}, \quad \delta(n,i) := t_i^n - t_{i-1}^n,\ i \geq 1. \tag{3}$$

Here, n_t is the stochastic number of observations recorded on the interval $[0,t]$ for $t \in \mathbb{R}_+$, while $\delta(n,i)$ is the i^{th} spacing of the observation times. For any process V, we denote $V_i^n := V_{t_i^n}$.

Let $\{\delta_n\}_n$ be a positive sequence of real numbers satisfying $\delta_n \to 0$ as $n \to \infty$. We may think of δ_n as the average magnitude of the spacings between successive observation times. If the observation times were equally spaced (the regular observation scheme), then δ_n would be proportional to that spacing. The difference between the regular observation scheme and the general scheme is characterized by two semimartingales $\alpha, \overline{\alpha}$, which are approximately the conditional mean and variance of the time differences, and characterizing the density of the observations. Specifically, conditional upon an appropriate σ-algebra, the expectations of $\delta(n,i)/\delta_n$ and $(\delta(n,i)\alpha_{i-1}^n - \delta_n)^2/\delta_n^2$ are approximately equal to $1/\alpha_{i-1}^n$ and $\overline{\alpha}_{i-1}^n$, respectively. For brevity, we move the details of the specifications to Assumption (O) in Appendix A. A useful consequence of our setting is the following convergence in probability:

$$\delta_n n_t \xrightarrow{\mathbb{P}} A_t := \int_0^t \alpha_s \mathrm{d}s. \tag{4}$$

The observation times framework is very general, and includes, e.g., *regular sampling scheme, time-changed regular sampling scheme, modulated Poisson sampling scheme, and predictably-modulated random walk sampling scheme,* see the discussion in Jacod et al. (2017).

2.3 Microstructure noise

The microstructure noise has a multiplicative form that allows for serial dependence, stochastic scale and dependence of the scale on the efficient price process. At time t_i^n, the microstructure noise is given by

$$\varepsilon_i^n := \gamma_i^n \cdot \chi_i.$$

Here, γ is a nonnegative Itô semimartingale on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$.[6] The process $\{\chi_i\}_{i\in\mathbb{Z}}$ is stationary and ρ-mixing, and its degree of serial dependence is controlled by a parameter v. Specifically, the autocovariance function of $\{\chi_i\}_i$ decays at a polynomial rate, i.e.,

$$|\mathbf{Cov}(\chi_i, \chi_{i+k})| \leq \frac{K}{k^v}, \tag{5}$$

where $K > 0$ is some positive constant. The reader is referred to Assumption (N) for the detailed specifications of γ and χ.

Remark 2.1. *To obtain limit results, we shall suppose that $v > 1$ for consistency and that $v > 2$ to derive the limit distribution, which allows for quite strong dependence close to the long memory boundary. Jacod et al. (2017) require $v > 0$ for consistency and $v > 1$ to establish the limit distribution.*

2.4 The observed noisy price

Finally, the observed noisy price Y_i^n is given by (for $i = 1, \ldots, n_t$)

$$Y_i^n = X_i^n + \varepsilon_i^n. \tag{6}$$

Note that both X and ε are latent, only Y is observable, and we aim to estimate the moments of ε using Y only.

3 The Design and the Intuition of the ReMeDI Estimators

3.1 The estimator of the autocovariance function

The intuition of the ReMeDI design can be best seen in a simpler setting. Let $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ be a stationary mixing sequence with mean zero; we would like to estimate its covariance

[6]The semimartingale assumption for γ is adopted for comparison with Jacod et al. (2017), but other assumptions can be accommodated and might be more reasonable.

$r_\ell := \mathbf{Cov}(\varepsilon_i, \varepsilon_{i+\ell})$. The natural estimator is the sample analogue

$$\widehat{r}_\ell := \frac{1}{n} \sum_{i=0}^{n-\ell} \varepsilon_i \varepsilon_{i+\ell}, \tag{7}$$

which is consistent and asymptotically normal under very mild conditions.

We consider instead an estimator that replaces the "observations" $\varepsilon_i, \varepsilon_{i+\ell}$ by the "differences", i.e.,

$$\widetilde{r}_\ell^n := \frac{1}{n} \sum_{i=k_n'}^{n-\ell-k_n} \left(\varepsilon_i - \varepsilon_{i-k_n'}\right) \left(\varepsilon_{i+\ell} - \varepsilon_{i+\ell+k_n}\right),$$

where k_n, k_n' are integers that grow at certain rates as the sample size increases. The estimator $\widetilde{r}_\ell^n$ follows the ReMeDI design and it provides another consistent estimator of r_ℓ, provided $k_n \wedge k_n' \to \infty$, and $\frac{k_n \vee k_n'}{n} \to 0$. The intuition of the consistency becomes immediate if one rewrites $\widetilde{r}_\ell^n$ as

$$\widetilde{r}_\ell^n = \frac{1}{n} \sum_{i=k_n'}^{n-\ell-k_n} \varepsilon_i \varepsilon_{i+\ell} - \frac{1}{n} \sum_{i=k_n'}^{n-\ell-k_n} \varepsilon_i \varepsilon_{i+\ell+k_n} - \frac{1}{n} \sum_{i=k_n'}^{n-\ell-k_n} \varepsilon_{i-k_n'} \varepsilon_{i+\ell} + \frac{1}{n} \sum_{i=k_n'}^{n-\ell-k_n} \varepsilon_{i-k_n'} \varepsilon_{i+\ell+k_n}. \tag{8}$$

The first average is (asymptotically) equivalent to the sample analogue, thus it converges in probability to r_ℓ; the remaining three averages are centred at $r_{\ell+k_n}$, $r_{\ell+k_n'}$, and $r_{\ell+k_n+k_n'}$, which themselves converge to zero as $n \to \infty$ at a rate depending on (5).

Taking differences seems redundant if the time series $\{\varepsilon_i\}_i$ is observable. However, in our framework, ε is masked by the efficient price X, and we only observe $Y = X + \varepsilon$. Taking time differences removes the effect of the efficient price. The intuition of such removal under infill asymptotics is that the differences of the efficient prices, say, $X_i^n - X_{i-k_n'}^n$, are much smaller than the differences of the noise as n increases.

3.2 The general ReMeDI design

We next formally define our ReMeDI (Realized moMents of Disjoint Increments) estimator of a general class of parameters. First, we provide some notations that we will use below. Let $\mathfrak{J}$ be the set of all finite sequences of integers satisfying

$$\mathfrak{J} := \left\{ \mathbf{j} = (j_1, j_2, \ldots, j_q) : j_l \in \mathbb{Z}, l = 1, 2, \ldots, q; q \geq 2 \right\}.$$

In the sequel, we will assume without loss of generality that $j_1 = \max\{j_l : j_l \in \mathbf{j}\}$ for any $\mathbf{j} \in \mathfrak{J}$. The $\mathbf{j}$-moments of χ, the stationary component of microstructure noise, are given by

$$r(\mathbf{j}) := \mathbb{E}\left(\prod_{l=1}^{q} \chi_{j_l}\right). \tag{9}$$

This is our parameter of interest (after proper scaling by the γ process); it includes the autocovariance function of the noise process and many other examples as special cases.

Let $\boldsymbol{k} = (k_1, \ldots, k_q)$ be a q-tuple of integers. For any $\mathbf{j} \in \mathfrak{J}$ and any process V, let $\mathbb{I}(\boldsymbol{k}, \mathbf{j})_t^n$ be the set of observation indices on $[0, t]$ for which the following *multi-difference operator* $\Delta_{\mathbf{j}}^{\boldsymbol{k}}(\cdot)_i^n$ is well defined[7]:

$$\Delta_{\mathbf{j}}^{\boldsymbol{k}}(V)_i^n := \prod_{l=1}^{q} \left(V_{i+j_l}^n - V_{i+j_l-k_l}^n\right). \tag{10}$$

Then the ReMeDI estimator corresponding to $r(\mathbf{j})$ based on data $\{Y_i^n\}_{i=1}^{n_t}$ and tuning parameters $\boldsymbol{k}$ is defined by

$$\text{ReMeDI}(Y; \mathbf{j}, \boldsymbol{k})_t^n := \sum_{i \in \mathbb{I}(\boldsymbol{k}, \mathbf{j})_t^n} \Delta_{\mathbf{j}}^{\boldsymbol{k}}(Y)_i^n. \tag{11}$$

Remark 3.1. *Using the above notations, we rewrite the estimator $\widetilde{r}_\ell^n$ as follows*

$$\widetilde{r}_\ell^n = \frac{1}{n} \sum_{i=k_n'}^{n-\ell-k_n} \Delta_{0,\ell}^{-k_n, k_n'}(\varepsilon)_i^n.$$

The general ReMeDI *approach inherits two salient features of this estimator determined by the choices of $\boldsymbol{k}$: 1) the first entry of $\boldsymbol{k}$ will be negative whereas the remaining ones are positive, i.e., the first difference is a* forward difference *and the remaining ones are* backward differences*; 2) $\forall 1 \leq l \leq q$, $|k_l| \to \infty$ as $n \to \infty$, and we will often write $\boldsymbol{k}_n = (k_{1,n}, \ldots, k_{q,n})$ in the sequel to reflect such dependence.*

We discuss a little more why the general ReMeDI procedure works under infill asymptotics. For this purpose, suppose that the noise size process γ is constant and we are estimating $\mathbb{E}\left(\prod_{l=1}^{q} \varepsilon_{i+j_l}^n\right)$. Suppose that $\boldsymbol{k}_n$ satisfies the two properties in Remark 3.1. Next, we explain how to connect $\mathbb{E}\left(\prod_{l=1}^{q} \varepsilon_{i+j_l}^n\right)$ and $\Delta_{\mathbf{j}}^{\boldsymbol{k}_n}(Y)_i^n$ with $\Delta_{\mathbf{j}}^{\boldsymbol{k}_n}(\varepsilon)_i^n$. To see this, we first note that $\{i + j_l - k_{l,n}\}_l$ are the "distant" indices of the intervals on which the backward and forward differences are taken. Figure 1 illustrates a simple example with $\mathbf{j} = (j_1, j_2, j_3)$, $\boldsymbol{k}_n = (-k_n, 2k_n, 4k_n)$ for some $k_n \in \mathbb{N}_+$. The forward difference starts at the $(i + j_1)$-th observation and ends at the $(i + j_1 + k_n)$-th observation; for

[7] By convention we set $\Delta_{\mathbf{j}}^{\boldsymbol{k}}(V)_i^n = 1$ if $\mathbf{j} = \emptyset$ and $\Delta_{\mathbf{j}}^{\boldsymbol{k}}(V)_i^n = 0$ if $\mathbf{j}$ is a singleton.

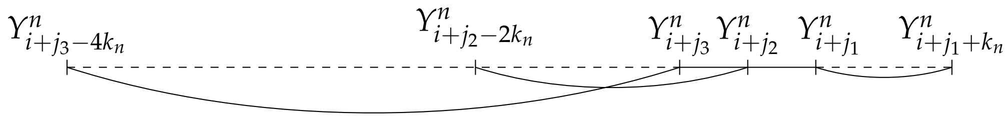

Figure 1: Illustration the ReMeDI estimator of $\mathbf{j}$-moments with $\mathbf{j} = (j_1, j_2, j_3)$ and $\boldsymbol{k}_n = (-k_n, 2k_n, 4k_n)$.

the remaining indices in $\mathbf{j}$, the associated differences start from $i + j_2, i + j_3$ and end at $i + j_2 - 2k_n, i + j_3 - 4k_n$, respectively. The intuition of the ReMeDI approach is that the "distant" noise terms are approximately independent of each other, and are also independent of the "clustered" noise $\{\varepsilon^n_{i+j_l}\}_l$ (recall a special case outlined in (8)), therefore

$$\mathbb{E}\left(\Delta^{\boldsymbol{k}_n}_{\mathbf{j}}(\varepsilon)^n_i\right) \approx \mathbb{E}\left(\prod_{l=1}^{q} \varepsilon^n_{i+j_l}\right).$$

If $k_{l,n}$ is still relatively small such that $\sup_l \delta_n |k_{l,n}| \to 0$, the differences/increments of the efficient price over the intervals are asymptotically negligible. That is, $\Delta^{\boldsymbol{k}_n}_{\mathbf{j}}(Y)^n_i \approx \Delta^{\boldsymbol{k}_n}_{\mathbf{j}}(\varepsilon)^n_i$. Thus the averages of $\Delta^{\boldsymbol{k}_n}_{\mathbf{j}}(Y)^n_i$ will converge in probability to $\mathbb{E}\left(\prod_{l=1}^{q} \varepsilon^n_{i+j_l}\right)$ by the law of large numbers. This is the intuition of the identification.

4 The Asymptotic Properties of the ReMeDI Estimators

4.1 Consistency

We next give the large sample properties of the ReMeDI estimator (for a given choice of $\boldsymbol{k}_n$) in our general setting. For a general γ process that satisfies Assumption (N), the "average size" of the noise moments $\mathbb{E}\left(\prod_{l=1}^{q} \varepsilon^n_{i+j_l}\right)$ is $\int_0^t \gamma_s^q \mathrm{d}A_s / A_t$, and this scaling appears in the probability limit of the ReMeDI estimators. Also recall (5) that v is the parameter that controls the degree of serial dependence in the noise.

Theorem 4.1. *Let Assumptions (H, O, N) hold, assume $v > 1$ and $\boldsymbol{k}_n$ satisfies*

$$\begin{cases} -k_{1,n} \to \infty, k_{l,n} \to \infty, \ \forall\, l \geq 2, \\ \sup_l \left|\delta_n^{\eta} k_{l,n}\right| \to 0, \ \eta \in (0, 1/2), \ \forall\, l \geq 1, \\ k_{l+1,n} - k_{l,n} \to \infty, \forall\, l \geq 2, \end{cases} \tag{12}$$

as $n \to \infty$. For $\mathbf{j} \in \mathfrak{J}$, we have the following convergence in probability:

$$\frac{\mathrm{ReMeDI}(Y; \mathbf{j}, \boldsymbol{k}_n)^n_t}{n_t} \xrightarrow{\mathbb{P}} \mathbf{R}(\mathbf{j})_t := \frac{\int_0^t \gamma_s^q \mathrm{d}A_s}{A_t} \boldsymbol{r}(\mathbf{j}), \tag{13}$$

where $\boldsymbol{r}(\mathbf{j})$ is defined in (9) and A_t in (4).

This says that our estimator consistently estimates $\boldsymbol{r}(\mathbf{j})$ up to a time t-varying scaling factor that depends on the average scale of the noise and on the stochastic process governing observation times.

Let $\{k_n\}_n$ be a sequence of integers satisfying $k_n \to \infty$, $k_n\delta_n \to 0$. Let $\boldsymbol{k}_n$ be specified as follows: $k_{l,n} = -k_n$ if $l = 1$, and $k_{l,n} = (l-1)k_n$ if $l \geq 2$. Then, $\boldsymbol{k}_n$ satisfies the conditions in (12).

4.2 Limit distribution

We first restrict further the values of $\boldsymbol{k}_n$ in order to facilitate the limit theory.[8] Among many possibilities, we propose the following specification of $\boldsymbol{k}_n$, which is solely determined by a single integer k_n:

$$k_{l,n} = \begin{cases} -k_n & \text{if } l = 1, \\ 2^{l-1}k_n & \text{if } l \geq 2, \end{cases} \tag{14}$$

where k_n is related to v as follows:

$$v > 2, \quad k_n\delta_n^{\eta} \to 0, \quad \eta \in \left(\frac{1}{2v}, \frac{1}{3}\right).$$

Remark 4.1. *Note that* (14) *implies* (12). *In the sequel, we will omit $\boldsymbol{k}_n$ and simply write $\Delta_{\mathbf{j}}(Y)_i^n$ and $\mathrm{ReMeDI}(Y;\mathbf{j})_t^n$ instead of $\Delta_{\mathbf{j}}^{\boldsymbol{k}_n}(Y)_i^n$ and $\mathrm{ReMeDI}(Y;\mathbf{j},\boldsymbol{k}_n)_t^n$ when $\boldsymbol{k}_n$ satisfies* (14).

We establish the CLT for both the following centered stochastic processes:

$$Z(\mathbf{j})_t^n := \frac{1}{\sqrt{\delta_n}}\left(\delta_n \mathrm{ReMeDI}(Y;\mathbf{j})_t^n - \boldsymbol{r}(\mathbf{j})\int_0^t \gamma_s^q \mathrm{d}A_s\right); \; \overline{Z}(\mathbf{j})_t^n := \sqrt{n_t}\left(\frac{\mathrm{ReMeDI}(Y;\mathbf{j})_t^n}{n_t} - \mathbf{R}(\mathbf{j})_t\right).$$

The first process involves unknown but deterministic norming, whereas the second process is normed by the observed stochastic sample size. Thus the second one is "feasible" in practice. Now let $\mathcal{F}_\infty := \bigvee_{t>0}\mathcal{F}_t$.

Theorem 4.2. *Let Assumptions (H), (O) and (N) hold, and k_n, v satisfy* (14). *For any $t > 0, \mathbf{j}, \mathbf{j}' \in \mathfrak{J}$, we have the following $\mathcal{F}_\infty$-stable convergence in law*

(a) $(Z(\mathbf{j})_t^n, Z(\mathbf{j}')_t^n) \xrightarrow{\mathcal{L}_s - \mathcal{F}_\infty} (Z(\mathbf{j})_t, Z(\mathbf{j}')_t)$ *, where the limit is defined on an extension $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ of $(\Omega, \mathcal{F}, \mathbb{P})$. Conditionally on $\mathcal{F}$, $(Z(\mathbf{j})_t, Z(\mathbf{j}')_t)$ are centred Gaussian with (co)variances $\sigma(\mathbf{j},\mathbf{j}')_t$ that is given by*

$$\sigma(\mathbf{j},\mathbf{j}')_t := \boldsymbol{s}(\mathbf{j},\mathbf{j}')\int_0^t \gamma_s^{q+q'}\mathrm{d}A_s + \boldsymbol{r}(\mathbf{j})\boldsymbol{r}(\mathbf{j}')\int_0^t \gamma_s^{q+q'}\overline{\alpha}_s \mathrm{d}A_s, \tag{15}$$

[8] In the supplementary material Li and Linton (2020), we discuss how to select k_n in practice.

where $\mathbf{s}(\mathbf{j},\mathbf{j}')$ *is given by* (B.6).

(b) $\left(\overline{Z}(\mathbf{j})^n_t, \overline{Z}(\mathbf{j}')^n_t\right) \xrightarrow{\mathcal{L}_s-\mathcal{F}_\infty} \left(\overline{Z}(\mathbf{j})_t, \overline{Z}(\mathbf{j}')_t\right)$*, where the limit is defined on an extension* $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ *of* $(\Omega, \mathcal{F}, \mathbb{P})$*. Conditionally on* $\mathcal{F}$*,* $\left(\overline{Z}(\mathbf{j})_t, \overline{Z}(\mathbf{j}')_t\right)$ *are centred Gaussian with (co)variances* $\overline{\sigma}(\mathbf{j},\mathbf{j}')_t$ *that is given by*

$$\overline{\sigma}(\mathbf{j},\mathbf{j}')_t := \frac{\mathbf{s}(\mathbf{j},\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q+q'} \mathrm{d}A_s + \frac{\mathbf{r}(\mathbf{j})\mathbf{r}(\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q+q'}\overline{\alpha}_s \mathrm{d}A_s + \frac{\mathbf{R}(\mathbf{j})_t\mathbf{R}(\mathbf{j}')_t}{A_t}\int_0^t \overline{\alpha}_s \mathrm{d}A_s - \frac{\mathbf{R}(\mathbf{j})_t\mathbf{r}(\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q'}\overline{\alpha}_s \mathrm{d}A_s - \frac{\mathbf{r}(\mathbf{j})\mathbf{R}(\mathbf{j}')_t}{A_t}\int_0^t \gamma_s^{q}\overline{\alpha}_s \mathrm{d}A_s. \tag{16}$$

Remark 4.2. $\mathbf{s}(\mathbf{j},\mathbf{j}')$ *is the asymptotic variance of the ReMeDI estimators contributed by the stationary part of the noise. The explicit form is given by* (B.6) *in Appendix B.1. When the observation scheme is simpler, e.g., when* $\overline{\alpha}_t \equiv 0$ *or* 1*, the asymptotic variance (as well as its consistent estimator) are much simplified, see the discussion in Appendix B.2.*

Remark 4.3 (Asymptotic variances of ReMeDI and LA)**.** *Note that the ReMeDI and LA estimators have very similar asymptotic (co)variances. The only difference lies in the* $\mathbf{s}(\mathbf{j},\mathbf{j}')$ *part, which represents the asymptotic variance contributed by the stationary part of noise. The* $\mathbf{s}(\mathbf{j},\mathbf{j}')$ *of the ReMeDI estimators includes the asymptotic (co)variances of the "distant" noise terms (recall the discussion in Section 3.2). It is therefore larger than the counterpart of the LA estimators. Hence, the LA estimators are asymptotically more efficient (although one can improve the efficiency of ReMeDI by taking averages of estimators computed using different* $\boldsymbol{k}_n$*). However, simulation studies show that the ReMeDI class works better in finite samples with realistic sample sizes (or equivalently, data frequency) — it has smaller finite sample variance and is almost unbiased under various model specifications. Moreover, the ReMeDI approach has greater computational efficiency, which pays off when one is working with massive high-frequency datasets (recall Footnote 4).*

Theorem 4.3. *Suppose that all the conditions of Theorem 4.2 hold. Moreover,* $\{i_n\}_n, \{\phi_n\}_n$ *are two sequences of integers satisfying:*

$$\frac{i_n}{k_n^v} \to 0, \quad i_n\delta_n^\eta \to 0, \quad \frac{\phi_n}{k_n\delta_n} \to 0, \quad \frac{k_n^{3/4}\delta_n}{\phi_n} \to 0. \tag{17}$$

For any $\mathbf{j} \in \mathfrak{J}$*, we have the following* $\mathcal{F}_\infty$*-stable convergence in law*

$$\frac{\sqrt{n_t}}{\sqrt{\widehat{\sigma}(\mathbf{j},\mathbf{j}')^n_t}}\left(\frac{\mathrm{ReMeDI}(Y;\mathbf{j})^n_t}{n_t} - \mathbf{R}(\mathbf{j})_t\right) \xrightarrow{\mathcal{L}_s-\mathcal{F}_\infty} \Phi, \tag{18}$$

where Φ is a standard normal random variable that is defined on an extension of the space and is independent of $\mathcal{F}$, and $\widehat{\sigma}(\mathbf{j},\mathbf{j}')_t^n$ is a consistent estimator of the asymptotic variance constructed in (B.7).

4.3 Estimating the autocovariances of microstructure noise

In this section we consider the special case concerning the estimation of the autocovariance function of the microstructure noise. Let $\mathbf{j}_\ell = (0,\ell), \ell \in \mathbb{N}_+$, and let

$$\widehat{R}^n_{t,\ell} := \frac{1}{n_t}\text{ReMeDI}(Y;\mathbf{j}_\ell)^n_t = \frac{1}{n_t}\sum_{i=2k_n}^{n_t-k_n-\ell}\left(Y^n_{i+\ell} - Y^n_{i+\ell+k_n}\right)\left(Y^n_i - Y^n_{i-2k_n}\right). \tag{19}$$

The following corollary provides the limit distribution.

Corollary 4.1 (ReMeDI estimators of autocovariances)**.** *Under the conditions of Theorem 4.2, we have*

$$\sqrt{n_t}\left(\widehat{R}^n_{t,\ell} - R_{t,\ell}\right) \overset{\mathcal{L}_s-\mathcal{F}_\infty}{\longrightarrow} \mathcal{N}\left(0, \overline{\sigma}(\mathbf{j}_\ell,\mathbf{j}_\ell)_t\right),$$

where

$$\overline{\sigma}(\mathbf{j}_\ell,\mathbf{j}_\ell)_t := \frac{1}{A_t}\left(\mathbf{s}_\ell\int_0^t \gamma_s^4 \mathrm{d}A_s + r_\ell^2\int_0^t \gamma_s^4\overline{\alpha}_s\mathrm{d}A_s + R^2_{t,\ell}\int_0^t \overline{\alpha}_s\mathrm{d}A_s - 2R_{t,\ell}r_\ell\int_0^t \gamma_s^2\overline{\alpha}_s\mathrm{d}A_s\right); \tag{20}$$

$$R_{t,\ell} := r_\ell\frac{\int_0^t \gamma_s^2\mathrm{d}A_s}{A_t}, \quad \mathbf{s}_\ell := \sum_{k=-\infty}^{\infty}\left(\mathbb{E}((\chi_0\chi_\ell - r_\ell)(\chi_k\chi_{k+\ell} - r_\ell)) + 3r_k^2\right).$$

Moreover, under the assumptions of Theorem 4.3, we have

$$\frac{\sqrt{n_t}}{\sqrt{\widehat{\sigma}(\mathbf{j}_\ell,\mathbf{j}_\ell)^n_t}}\left(\widehat{R}^n_{t,\ell} - R_{t,\ell}\right) \overset{\mathcal{L}_s-\mathcal{F}_\infty}{\longrightarrow} \Phi, \tag{21}$$

where Φ is a standard normal random variables as in Theorem 4.3 and $\widehat{\sigma}(\mathbf{j}_\ell,\mathbf{j}_\ell)^n_t$ is provided in (B.7).

Remark 4.4. *$\mathbf{s}_\ell$ represents the variance of the ReMeDI estimators contributed by the stationary part of noise. It has two components. The first part $\sum_{k=-\infty}^{\infty}\mathbb{E}((\chi_0\chi_\ell - r_\ell)(\chi_k\chi_{k+\ell} - r_\ell))$ is in fact the asymptotic variance of the sample analogue, recall* (7)*. The second part $3\sum_{k=-\infty}^{\infty} r_k^2$ is the asymptotic variance of the three additional terms appear in* (8) *that arise in differencing.*

Remark 4.5. *The last three terms of $\overline{\sigma}(\mathbf{j}_\ell,\mathbf{j}_\ell)_t$ that appear in* (20) *arise because of the stochastic sampling scheme; it is positive and is zero whenever $r_\ell = 0$, $\overline{\alpha}_s \equiv 0$ or $\gamma_s \equiv K$, where K is a constant.*

We note that while the multiplicative structure of microstructure noise (recall (A.3)) allows for a time-varying and stochastic size of the noise, the serial correlation of the noise is not affected by the size process. This structure allows us to estimate the autocorrelations of noise directly once we have an estimator of the autocovariances. Define the ReMeDI estimator of the noise autocorrelation, $\widehat{r}(\ell)_t^n := \widehat{R}_{t,\ell}^n / \widehat{R}_{t,0}^n$, and its asymptotic variance estimator

$$\widehat{\sigma}(\ell)_t^n := \frac{\widehat{\sigma}(\mathbf{j}_\ell, \mathbf{j}_\ell)_t^n n_t^2}{(\text{ReMeDI}(Y, \mathbf{j}_0)_t^n)^2} - \frac{2n_t^2 \widehat{\sigma}(\mathbf{j}_0, \mathbf{j}_\ell)_t^n \text{ReMeDI}(Y, \mathbf{j}_\ell)_t^n}{(\text{ReMeDI}(Y, \mathbf{j}_0)_t^n)^3} + \frac{n_t^2 (\text{ReMeDI}(Y, \mathbf{j}_\ell)_t^n)^2 \widehat{\sigma}(\mathbf{j}_0, \mathbf{j}_0)_t^n}{(\text{ReMeDI}(Y, \mathbf{j}_0)_t^n)^4}.$$

The following corollary spells out the limit distribution of the proposed estimators.

Corollary 4.2 (ReMeDI estimators of autocorrelations). *Under the conditions of Theorem 4.3, we have the following $\mathcal{F}_\infty$-stable convergence in law*

$$\sqrt{\frac{n_t}{\widehat{\sigma}(\ell)_t^n}} \left(\widehat{r}(\ell)_t^n - r(\ell)\right) \xrightarrow{\mathcal{L}_s - \mathcal{F}_\infty} \Phi,$$

where Φ is a standard normal random variable as in Theorem 4.3.

5 Simulation Study

5.1 Model settings

We suppose that the efficient price process has stochastic volatility and jumps that appear in both the price and volatility processes:

$$\begin{aligned} &\mathrm{d}X_t = \kappa_1(\mu_1 - X_t)\mathrm{d}t + \sigma_t \mathrm{d}W_{1,t} + \xi_{1,t}\mathrm{d}N_t; \quad \mathrm{d}\sigma_t^2 = \kappa_2(\mu_2 - \sigma_t^2)\mathrm{d}t + \eta\sigma_t \mathrm{d}W_{2,t} + \xi_{2,t}\mathrm{d}N_t; \\ &\mathbf{Corr}(W_1, W_2) = v; \quad \xi_{1,t} \sim \mathcal{N}(0, \mu_2/10); \quad N_t \sim \text{Poi}(\lambda); \quad \xi_{2,t} \sim \text{Exp}(\delta). \end{aligned} \tag{22}$$

We set

$$\kappa_1 = 0.5;\ \mu_1 = 3.6;\ \kappa_2 = 5/252;\ \mu_2 = 0.04/252;\ \eta = 0.05/252;\ v = -0.5;\ \lambda = 1;\ \delta = \eta.$$

This setting is motivated by some empirical facts that jumps in price levels and volatility tend to occur together, see Todorov and Tauchen (2011).

We further suppose that the stationary component of the microstructure noise follows an AR(1) process with Gaussian innovations

$$\chi_{i+1} = \rho\chi_i + e_i, \quad e_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, 1-\rho^2\right), \quad |\rho| < 1.$$

Note that χ has unit variance. We set $\rho = 0.7$, motivated by the empirical studies in Aït-Sahalia et al. (2011) and Li et al. (2020).

5.2 LA versus ReMeDI

We estimate the autocovariances of microstructure noise using the ReMeDI estimator and the local averaging (LA) estimators (Jacod et al., 2017). We assume that the noise is stationary so that we can compare the estimates to the true parameters. We also assume that the observation scheme is regular so that we know explicitly the data frequency, which is a key factor that affects the finite sample performance of many high-frequency estimators.

The top and middle panels of Figure 2 present the estimation of the first 20 autocovariances of the noise by ReMeDI and LA.[9] The solid lines are the mean estimates over 1,000 replications; the shaded region represents the 95% simulated confidence intervals. We simulate 23,400 observations for each sample path, corresponding to the number of seconds in a business day (6.5 trading hours). The ReMeDI estimators perform well: the estimates are approximately unbiased with narrow confidence bands. Surprisingly, there is a significant average deviation of the LA estimates from the true parameters, and the confidence bands are much larger as well.

The deviation of the LA estimates is elicited by a *finite sample bias*, which is known to be a fraction of the *prior unknown* quadratic variation (QV) of the efficient price, see the discussion in Jacod et al. (2017). Thus to correct the bias, we need an estimate of the QV. But the estimation of QV in the presence of dependent noise is not trivial, see a discussion in Li et al. (2020). In a simulation context, we can obtain the QV and thus can give the LA estimators the privilege to make the bias correction, which is, of course not feasible in practice. The bottom panel of Figure 2 displays the bias corrected estimation of LA. Even with accurate bias correction, however, the ReMeDI estimators still outperform the LA estimators with almost no bias but greater accuracy.

It is interesting to compare ReMeDI and LA when the data frequencies vary. However, increasing the data frequency in a fixed time span has two effects: both the number of observations and the noise-to-signal ratio of tick returns will increase. We design a simulation study to separate the two effects and examine how sensitive ReMeDI and LA are to these changes.

The left panel of Figure 3 presents the mean-squared-error (MSE) of the ReMeDI and LA estimators for the first 20 autocovariances of the noise. The sample size varies from 23,400 (1 trading day) to 117,000 (1 trading week), and 468,000 (1 trading

[9]We select the same tuning parameter for the LA estimator as in Jacod et al. (2017); we also check other alternatives, and we find $k_n = 6$ leads to smaller bias.

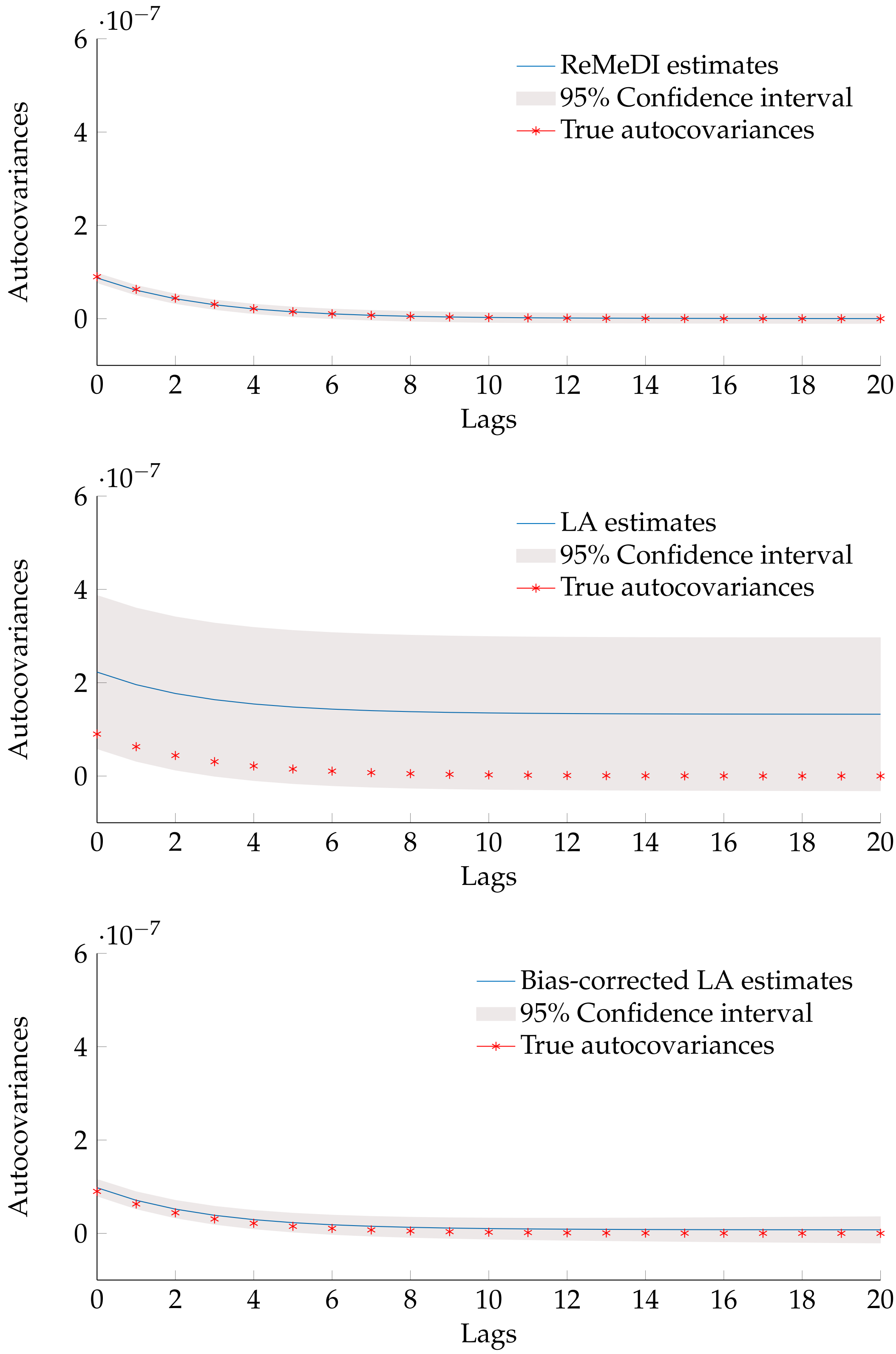

Figure 2: Estimation of the autocovariances of noise by the ReMeDI method (top panel), the local averaging method (middle panel) and the bias corrected local averaging method (bottom panel). The blue solid line is the mean estimates of 1,000 simulations by the three estimators. The tuning parameters of the ReMeDI and LA estimators are 10 and 6, respectively. The noise scale is fixed at $\gamma \equiv 5 \times 10^{-4}$.

month). The MSE of the ReMeDI estimators remains low and slightly drops when the sample sizes increases. The LA estimators, however, has larger MSE in a larger sample! This is statistically counterintuitive. However, it does make sense if we recall that the integrated volatility contributes to the finite sample bias of the LA estimators. Hence longer time span induces larger integrated volatility (relatively to the number of observations), which in turn leads to a larger finite sample bias. This is especially so if the sample covers a period of volatility burst, and the likelihood of an such event increases if the sampling period becomes large, see our empirical studies with real transaction prices.

The right panel of Figure 3 compares ReMeDI and LA when noise variance varies from 10^{-8} (small noise) to 10^{-6} (large noise). We note that the advantage of ReMeDI over LA is more prominent when the noise is smaller. Indeed, the size of noise in practice is closer to the small noise scenario, see an extensive empirical study by Christensen et al. (2014). Thus in an extreme case when the noise has *identical* statistical properties in two samples, LA may give very different estimates due to the differences in sample sizes or noise-to-signal ratios. The ReMeDI approach remains robust and accurate.

5.3 Random noise size and observation times

As the last robustness check, we now allow for stochastic observation times and random scales of noise. Following Jacod et al. (2017), we let $\{t_i^n\}$ follow an inhomogeneous Poisson process with rate $n\alpha_t$ where $\alpha_t = (1 + \cos(2\pi t))/2$ and the process γ satisfies

$$\gamma_t = C_\gamma \gamma'_t, \quad \mathrm{d}\gamma'_t = -\rho_\gamma(\gamma'_t - \mu_t)\mathrm{d}t + \sigma_\gamma \mathrm{d}W_t.$$

We set $\rho_\gamma = 10$, $\mu_t = 1 + 0.1\cos(2\pi t)$, $\sigma_\gamma = 0.1$, $C_\gamma = 5 \times 10^{-4}$. Figure 4 reports the estimation of the autocorrelation functions by the two estimators. We observe similar patterns presented in Figure 2: compared to the ReMeDI estimators, the LA estimators have large biases with wide confidence band.

The supplementary material Li and Linton (2020) provides additional simulation studies to examine the CLT, the effect of rounding error due to the discreteness of price and sensitivity to the choice of tuning parameters.

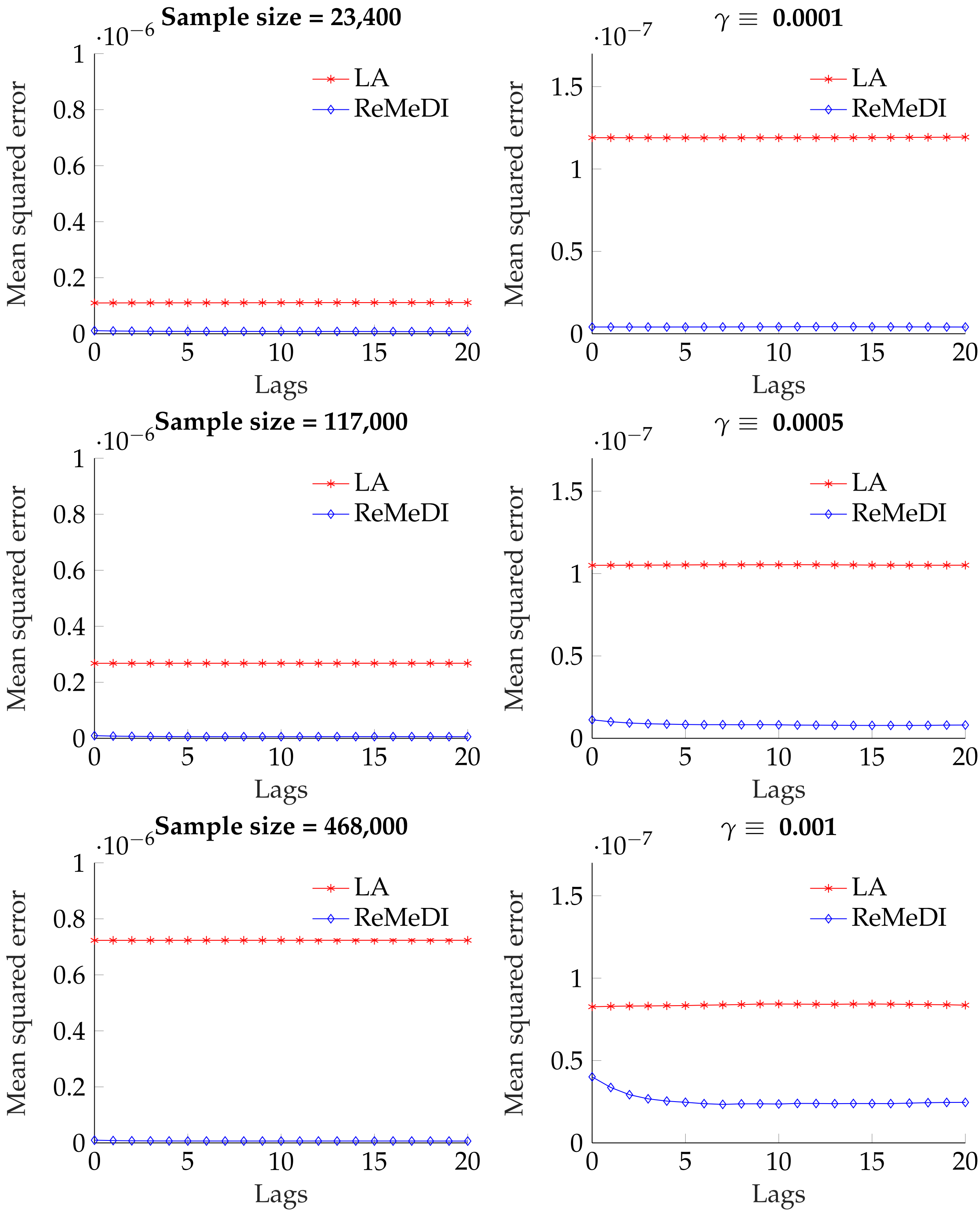

Figure 3: Mean squared error (MSE) of the ReMeDI and LA estimators for the first 20 autocovariances of noise based on 1,000 simulations. In the left panel, the noise scale is fixed at $\gamma = 5 \times 10^{-4}$ and the sample size varies; in the right panel, the size sample is 23,400 while the noise scale parameter varies. The tuning parameters of the ReMeDI and LA estimators are 10 and 6, respectively.

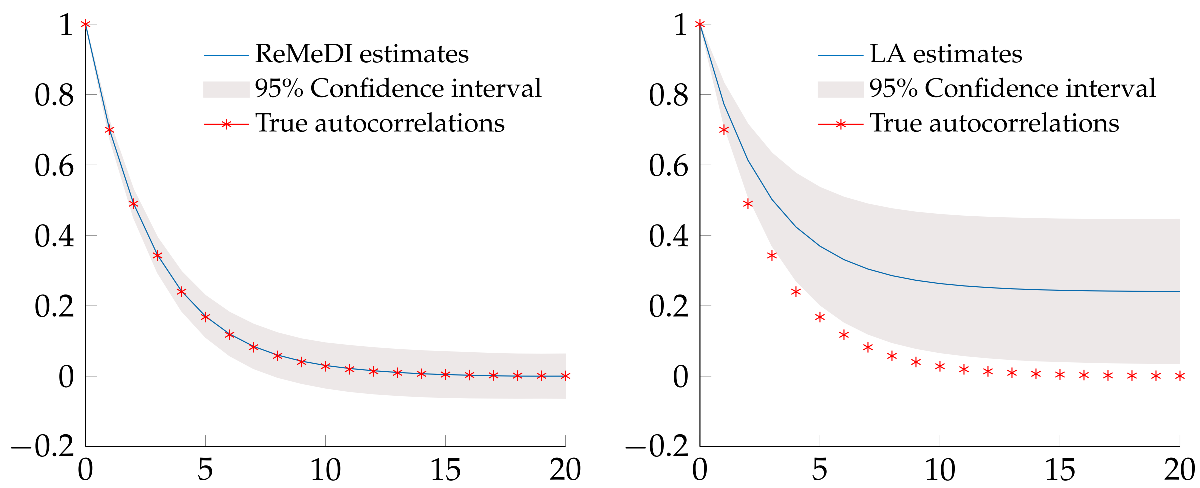

Figure 4: Estimation of the autocorrelations of noise by the ReMeDI method (left panel) and the local averaging method (right panel). The blue solid line is the mean estimates of 1,000 simulations by the two estimators. The tuning parameters of the ReMeDI and LA estimators are 10 and 6, respectively. The noise has stochastic scales and the observation times are random, see the specifications in Section 5.3.

6 Empirical Study

We obtain the transaction prices of Coca-Cola (trading symbol KO)[10] from the TAQ database for January 2018 (21 trading days). We remove prices before 9:30 and after 16:00. We collect approximately 50,000 observations per day, i.e., 2.1 transactions per second on average. The average price is 46.84$, with a standard deviation of 0.85.

Figure 5 plots the estimated autocovariances of noise by the ReMeDI estimators (the blue plots) based on samples of different sizes. The autocorrelation pattern is non-trivial: noise exhibits positive autocorrelations up to 4 lags and shortly thereafter, the sign switches to negative for a few lags, and then reverts to positive autocorrelations before decaying to zero around 20 lags. The pointwise confidence interval[11] includes zero or excludes positive values after lag 5, which is incompatible with simple long memory.

The ReMeDI estimates of microstructure noise presented in Figure 5 are economically intuitive. The positive autocovariances at the first several lags may be a consequence of the order splitting strategies by high-frequency traders (Biais et al. (1995)), or the successive transactions executed by limit orders (Parlour (1998)).[12] The negative

[10]In the supplementary material Li and Linton (2020), we use the transaction prices of General Electric (GE) and Citigroup (Citi), and we obtain similar results.

[11] Recall Section B.2 that the duration of successive observed prices is part of the asymptotic variance estimator. We do not plot the confidence intervals when we use transaction prices on different trading days since the prices will cover overnight non-trading hours.

[12]Hasbrouck and Ho (1987) and Choi et al. (1988) model the continuation of order flows by an AR(1)

autocovariances at the intermediate lags are consistent with the prediction of inventory models (Ho and Stoll (1981), Hendershott and Menkveld (2014)), in which the market makers induce negatively autocorrelated order flows to balance his inventories. However, the LA method gives very different estimates: it says that the noise is strongly autocorrelated without any sign of decay after 20 lags. This is economically counterintuitive — such a pattern, if it exists, would be exploited by high-frequency traders and we would expect it to disappear rapidly. Moreover, the serial dependence, according to the LA estimates, is even stronger when estimation is performed in a larger sample. Since we only estimate autocovariances of noise up to 20 ticks/lags, or a few seconds, it is statistically counterintuitive to obtain stronger autocovariance estimates using the prices of a week than using the prices in a single trading day. This is in line with our simulation study that the LA estimates are subject to a finite sample bias that depends on the noise-to-signal ratio and sample size. The ReMeDI approach retains its accuracy and robustness.

7 Concluding Remarks

We introduced a nonparametric method to separate the microstructure noise from the underlying semimartingale efficient prices in a general setting. We demonstrate the robustness of the proposed method compared to the main existing approach. We have concentrated on the infill setting primarily and the univariate case. The method naturally extends to the multivariate case, although in that case, several issues arise. First, the nonsynchronous trading issue has to be faced. Second, even when the assets trade on a common clock, there are some remaining theoretical results that need to be established for the infill case. We have not discussed efficiency in a great deal, but one can improve efficiency in two ways: first, by combining the estimators associated with different choices of $\boldsymbol{k}_n$ by a minimum distance, and second by doing a kind of GLS procedure using a local estimator of γ_u. We leave these problems for future research.

Appendix A Assumptions and Regularity Conditions

Assumption (H). *The process b is locally bounded, the process σ is càdlàg , there is a localizing sequence $\{\tau_n\}_n$ of stopping times and for each n a deterministic nonnegative function Γ_n on E satisfying $\int \Gamma_n^2(z)\lambda(\mathrm{d}z) < \infty$ such that $|\vartheta(\omega, t, z)| \wedge 1 \leq \Gamma_n(z)$ for all (ω, t, z) satisfying $t \leq \tau_n(\omega)$.*

process.

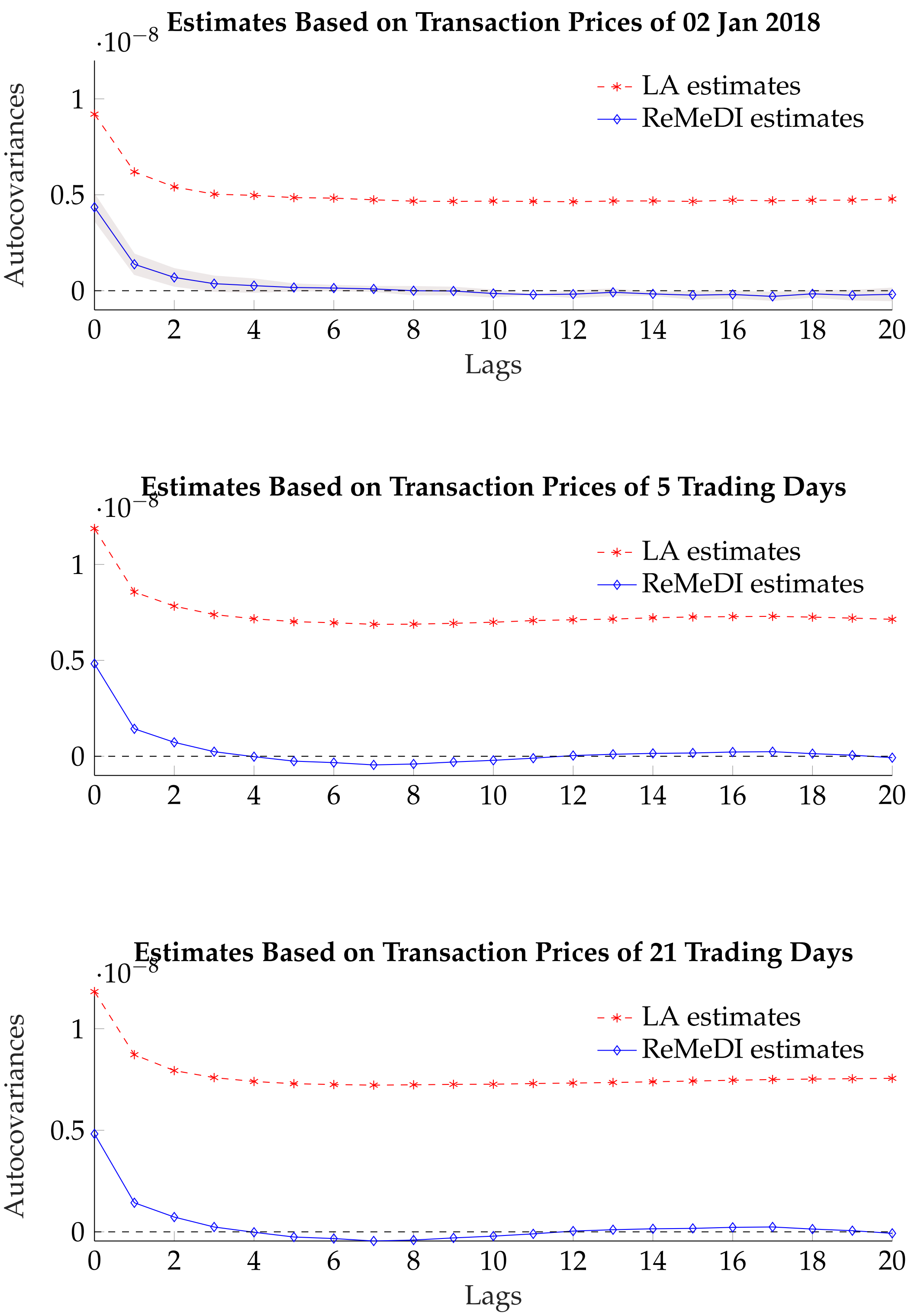

Figure 5: Estimation of autocovariances of noise for Coca-Cola (KO) in January 2018. In the top panel, we use the transaction prices of KO on 2 January 2018; in the middle panel, we use the transaction prices of KO in the second trading week (8 January 2018 to 12 January 2018); we employ the entire transaction prices of KO in January 2018 in the bottom panel. The tuning parameters for ReMeDI and LA are 10 and 6, respectively. The shaded area in the top panel represents the 95% confidence interval, and we set $i_n = 5, \phi_n = k_n^{3/5}/n$ to compute the asymptotic variances of the ReMeDI estimators, where n is the number of observations.

Definition A.1. *Let $\{\chi_i\}_{i\in\mathbb{Z}}$ be a sequence of stationary random variables. For any $k \in \mathbb{N}_+$, we define the following* mixing coefficients *for $k \in \mathbb{N}_+$:*

$$\rho_k := \sup\left\{|\mathbb{E}(V_h V_{k+h})| : \mathbb{E}(V_k) = \mathbb{E}(V_{k+h}) = 0, \|V_h\|_2 \le 1, \|V_{k+h}\|_2 \le 1, V_h \in \mathcal{G}_h, V_{k+h} \in \mathcal{G}^{k+h}\right\}, \tag{A.1}$$

where $\mathcal{G}_p := \sigma\{\chi_k : p \ge k\}$, $\mathcal{G}^q := \sigma\{\chi_k : q \le k\}$. The sequence $\{\chi_i\}_{i\in\mathbb{Z}}$ is ρ mixing *if $\rho_k \to 0$ as $k \to \infty$.*

Assumption (O). *$\alpha, \overline{\alpha}$ are two Itô semimartingales defined on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\ge0}, \mathbb{P})$ satisfying Assumption (H). We further assume there is a localizing sequence $\{\tau_m\}_m$ of stopping times and positive constants $\kappa_{m,p}$ and κ such that:*

(i) For $t < \tau_m$, we have $\frac{1}{\kappa_{m,1}} \le \alpha_{t-} \le \kappa_{m,1}$ and $\overline{\alpha}_{t-} \le \kappa_{m,1}$.

(ii) Let $(\mathcal{F}^n_t)_{t\ge0}$ be the smallest filtration satisfying

(a) $\mathcal{F}_t \subset \mathcal{F}^n_t$,

(b) t^n_i is a $\{\mathcal{F}^n_t\}_{t\ge0}$ stopping time for $i = 0, 1, 2, \ldots$,

(c) $\delta(n,i)$, conditional $\mathcal{F}^n_{i-1} := \mathcal{F}^n_{t^n_{i-1}}$, is independent of $\mathcal{F}_\infty := \bigvee_{t\ge0} \mathcal{F}_t$ for $i = 0, 1, 2, \ldots$

(iii) With the restriction $\{t^n_{i-1} < \tau_m\}$, and for all $p > 0$,

$$\begin{aligned}
&\left|\mathbb{E}\left(\delta(n,i)\,\middle|\,\mathcal{F}^n_{i-1}\right) - \frac{\delta_n}{\alpha^n_{i-1}}\right| \le \kappa_{m,1}\delta_n^{\frac{3}{2}+\kappa}, \\
&\left|\mathbb{E}\left(\left(\delta(n,i)\alpha^n_{i-1} - \delta_n\right)^2 \,\middle|\,\mathcal{F}^n_{i-1}\right) - \delta_n^2\overline{\alpha}^n_{i-1}\right| \le \kappa_{m,2}\delta_n^{2+\kappa}, \\
&\mathbb{E}\left(\delta(n,i)^p\,\middle|\,\mathcal{F}^n_{i-1}\right) \le \kappa_{m,p}\delta_n^p.
\end{aligned} \tag{A.2}$$

Assumption (N). *Let $\{\chi_i\}_{i\in\mathbb{Z}}$ be a stationary ρ-mixing random sequence with mixing coefficients $\{\rho_k\}_{k\in\mathbb{N}_+}$. At stage n, the noise at time t^n_i is given by*

$$\varepsilon^n_i = \gamma^n_i \cdot \chi_i, \tag{A.3}$$

where γ is a nonnegative Itô semimartingale on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\ge0}, \mathbb{P})$ satisfying Assumption (H) and is not identically zero on any interval. We further assume that $\{\chi_i\}_{i\in\mathbb{Z}}$ is centred at 0 with variance 1 and finite moments of all orders, and is independent of $\mathcal{F}_\infty$. Moreover, there is some $K > 0, v > 0$ such that

$$\rho_k \le \frac{K}{k^v} \quad \forall\, k \in \mathbb{N}_+. \tag{A.4}$$

Appendix B Asymptotic (Co)Variance and its Estimation

B.1 The asymptotic (co)variance

This section introduces $\boldsymbol{s}(\mathbf{j},\mathbf{j}')$ that appears in the asymptotic variance in Theorem 4.2. In the sequel whenever we have two vectors $\mathbf{j}=(j_1,\dots,j_q),\mathbf{j}'=(j'_1,\dots,j'_{q'})\in\mathfrak{J}$, we suppose without loss of generality that $q\le q'$. We denote

$$\begin{aligned}
&\mathbf{j}\oplus\mathbf{j}'=(j_1,j_2,\dots,j_q,j'_1,j'_2,\dots,j'_{q'}),\quad \mathbf{j}_{-l}=\mathbf{j}\backslash\{j_l\},\\
&\mathbf{j}(+k)=(j_1+k,j_2+k,\dots,j_q+k),\ \text{for}\ k\in\mathbb{Z},\\
&\mathbf{j}_{Q_q}=(j_l:l\in Q_q)\ \text{for}\ Q_q\subset\{1,2,\dots,q\},\\
&\mathcal{Q}_q:=\{Q_q:Q_q\subsetneq\{1,2,\dots,q\}\}.
\end{aligned}$$

For each $Q_q\subset\{1,2,\dots,q\}$, there is an associated (unique) pair of subsets:

$$Q_q^c:=\{1,2,\dots,q\}\backslash Q_q,\quad Q_{q'}:=Q_q\cup\{q+1,\dots,q'\}. \tag{B.5}$$

We denote for each $k\in\mathbb{Z}$ the following moments[13]

$$\begin{aligned}
\boldsymbol{s}_0(\mathbf{j},\mathbf{j}';k)&:=\boldsymbol{r}\left(\mathbf{j}\oplus\left(\mathbf{j}'(+k)\right)\right)-\boldsymbol{r}\left(\mathbf{j}\right)\boldsymbol{r}\left(\mathbf{j}'\right);\\
\boldsymbol{s}_1(\mathbf{j},\mathbf{j}';k)&:=\sum_{Q_q\in\mathcal{Q}_q}\boldsymbol{r}\left(\mathbf{j}_{Q_q}\oplus\left(\mathbf{j}'_{Q_{q'}}(+k)\right)\right)\prod_{l\in Q_q^c}\boldsymbol{r}(j_l,j'_l+k);\\
\boldsymbol{s}_2(\mathbf{j},\mathbf{j}';k)&:=\sum_{\substack{j_l\in\mathbf{j},j'_{l'}\in\mathbf{j}'\\ l\neq l'}}\boldsymbol{r}(j_l,j'_{l'}+k)\boldsymbol{r}(\mathbf{j}_{-l})\boldsymbol{r}(\mathbf{j}'_{-l'})-\sum_{j_l\in\mathbf{j}}\boldsymbol{r}(\{j_l\}\oplus\mathbf{j}'(+k))\boldsymbol{r}(\mathbf{j}_{-l})\\
&\quad-\sum_{j'_{l'}\in\mathbf{j}'}\boldsymbol{r}(\{j'_{l'}+k\}\oplus\mathbf{j})\boldsymbol{r}(\mathbf{j}'_{-l'});
\end{aligned}$$

Then $\boldsymbol{s}(\mathbf{j},\mathbf{j}')$ is given by

$$\boldsymbol{s}(\mathbf{j},\mathbf{j}'):=\sum_{k\in\mathbb{Z}}\boldsymbol{s}_0(\mathbf{j},\mathbf{j}';k)+\boldsymbol{s}_1(\mathbf{j},\mathbf{j}';k)+\boldsymbol{s}_2(\mathbf{j},\mathbf{j}';k). \tag{B.6}$$

B.2 The estimation of the asymptotic (co)variance

First, we introduce a sequence of notations

$$\widehat{\delta}_i^n:=\left(\frac{k_n\delta(n,i+1+k_n)-t^n_{i+2+2k_n}+t^n_{i+2+k_n}}{(t^n_{i+k_n}-t^n_i)\vee\phi_n}\right)^2,\quad U(1)^n_t:=\sum_{i=0}^{n_t-w(1)_n}\widehat{\delta}_i^n,$$

[13] By convention we let $\boldsymbol{r}(\emptyset)=1$.

$$U(2;\mathbf{j})^n_t := \sum_{i=0}^{n_t-w(2)_n} \widehat{\delta}^n_i \Delta_{\mathbf{j}}(Y)^n_{i+w(2)^n_2};$$

$$U(3,\mathbf{j},\mathbf{j}')^n_t := \sum_{i=0}^{n_t-w(3)_n} \widehat{\delta}^n_i \Delta_{\mathbf{j}}(Y)^n_{i+w(3)^n_2}\Delta_{\mathbf{j}'}(Y)^n_{i+w(3)^n_3},$$

$$U(4;\mathbf{j},\mathbf{j}')^n_t := -\sum_{i=2^{q-1}k_n}^{n_t-w(4)_n} \Delta_{\mathbf{j}}(Y)^n_i\Delta_{\mathbf{j}'}(Y)^n_{i+w(4)^n_2},$$

$$U(5,k;\mathbf{j},\mathbf{j}')^n_t := \sum_{Q_q\in\mathcal{Q}_q}\sum_{i=2^{e(Q_q)}k_n}^{n_t-w(5)_n} \Delta_{\mathbf{j}_{Q_q}\oplus(\mathbf{j}'_{Q_{q'}}(+k))}(Y)^n_i \prod_{\ell:l_\ell\in Q_q^c}\Delta_{(j_{l_\ell},j'_{l_\ell}+k)}(Y)^n_{i+w(5)^n_{\ell+1}},$$

$$U(6,k;\mathbf{j},\mathbf{j}')^n_t := \sum_{\substack{j_l\in\mathbf{j},j'_{l'}\in\mathbf{j}'\\ l\neq l'}}\sum_{i=2k_n}^{n_t-w(6)_n} \Delta_{(j_l,j'_{l'}+k)}(Y)^n_i\Delta_{\mathbf{j}_{-l}}(Y)^n_{i+w(6)^n_2}\Delta_{\mathbf{j}'_{-l'}}(Y)^n_{i+w(6)^n_3}$$

$$-\sum_{j_l\in\mathbf{j}}\sum_{i=2^{q'}k_n}^{n_t-w'(6)_n} \Delta_{\{j_l\}\oplus\mathbf{j}'(+k)}(Y)^n_i\Delta_{\mathbf{j}_{-l}}(Y)^n_{i+w'(6)^n_2}$$

$$-\sum_{j'_{l'}\in\mathbf{j}'}\sum_{i=2^{q}k_n}^{n_t-w''(6)_n} \Delta_{\{j'_{l'}+k\}\oplus\mathbf{j}}(Y)^n_i\Delta_{\mathbf{j}'_{-l'}}(Y)^n_{i+w''(6)^n_2},$$

$$U(7,k;\mathbf{j},\mathbf{j}')^n_t := \mathrm{ReMeDI}(\mathbf{j}\oplus\mathbf{j}'(+k))^n_t;\quad U(k;\mathbf{j},\mathbf{j}')^n_t := \sum_{\ell=5}^{7} U(\ell,k;\mathbf{j},\mathbf{j}')^n_t,$$

where the indices appear above are given by

$$w(1)_n := 2+2k_n,\ w(2)^n_2 := 2+(3+2^{q-1})k_n,\ w(2)_n := w(2)^n_2+j_1+k_n;$$
$$w(3)^n_2 := 2+(3+2^{q-1})k_n,\ w(3)^n_3 := 2+(5+2^{q-1}+2^{q'-1})k_n+j_1;$$
$$w(3)_n := w(3)^n_3+j'_1+k_n;\ w(4)^n_2 := 2k_n+q'_n+j_1,\ w(4)_n := w(4)^n_2+j'_1+k_n;$$
$$e(Q_q) := (2|Q_q|+q'-q-1)\vee 1,\quad w(5)^n_{\ell+1} := 4\ell k_n+\sum_{\ell'=1}^{\ell} j_{l_{\ell'}}\vee(j'_{l_{\ell'}}+k)\ \text{ for } \ell\geq 1,$$
$$w(5)_n := w(5)^n_{|Q_q^c|+1}+j_{l_{|Q_q^c|}}\vee(j'_{l_{|Q_q^c|}}+k)+k_n;$$
$$w(6)^n_2 := (2^{q-2}+2)k_n+j_\ell\vee(j'_{\ell'}+k),\ w(6)^n_3 := (2^{q-2}+2^{q'-2}+2)k_n+j_1+j_\ell\vee(j'_{\ell'}+k),$$
$$w'(6)^n_2 := (2^{q-2}+2)k_n+j_\ell\vee(j'_1+k),\ w''(6)^n_2 := (2^{q'-2}+1)k_n+(j'_{\ell'}+k)\vee j_1,$$
$$w(6)_n := w(6)^n_3+j'_1+k_n,\ w'(6)_n := w'(6)^n_2+j_1+k_n,\ w''(6)_n := w''(6)^n_2+j'_1+k_n.$$

The asymptotic variance estimator is given by

$$\widehat{\sigma}(\mathbf{j},\mathbf{j}')^n_t := \frac{1}{n_t}\sum_{\ell=1}^{3}\widehat{\sigma}_\ell(\mathbf{j},\mathbf{j}')^n_t, \tag{B.7}$$

where

$$
\begin{aligned}
\widehat{\sigma}_1(\mathbf{j},\mathbf{j}')_t^n &:= U(0;\mathbf{j},\mathbf{j}')_t^n + \sum_{k=1}^{i_n}\left(U(k;\mathbf{j},\mathbf{j}')_t^n + U(k;\mathbf{j}',\mathbf{j})_t^n\right) + (2i_n+1)U(4;\mathbf{j},\mathbf{j})_t^n;\\
\widehat{\sigma}_2(\mathbf{j},\mathbf{j}')_t^n &:= U(3,\mathbf{j},\mathbf{j}');\\
\widehat{\sigma}_3(\mathbf{j},\mathbf{j}')_t^n &:= \frac{1}{n_t^2}\mathrm{ReMeDI}(Y;\mathbf{j})_t^n\mathrm{ReMeDI}(Y;\mathbf{j}')_t^n U(1)_t^n\\
&\quad - \frac{1}{n_t}\left(\mathrm{ReMeDI}(Y;\mathbf{j})_t^n U(2,\mathbf{j}')_t^n + \mathrm{ReMeDI}(Y;\mathbf{j}')_t^n U(2,\mathbf{j})_t^n\right).
\end{aligned}
$$

The estimators seem quite complicated. However, the intuition will be clear in light of the following convergences, which are proved in the supplementary material Li and Linton (2020):

$$
\begin{aligned}
&\frac{1}{n_t}\widehat{\sigma}_1(\mathbf{j},\mathbf{j}')_t^n \xrightarrow{\mathbb{P}} \frac{\boldsymbol{s}(\mathbf{j},\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q+q'}\mathrm{d}A_s, \quad \frac{1}{n_t}\widehat{\sigma}_2(\mathbf{j},\mathbf{j}')_t^n \xrightarrow{\mathbb{P}} \frac{\boldsymbol{r}(\mathbf{j})\boldsymbol{r}(\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q+q'}\overline{\alpha}_s\mathrm{d}A_s;\\
&\frac{1}{n_t}\widehat{\sigma}_3(\mathbf{j},\mathbf{j}')_t^n \xrightarrow{\mathbb{P}} \frac{\mathbf{R}(\mathbf{j})\mathbf{R}(\mathbf{j}')}{A_t}\int_0^t \overline{\alpha}_s\mathrm{d}A_s - \frac{\mathbf{R}(\mathbf{j})\boldsymbol{r}(\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q'}\overline{\alpha}_s\mathrm{d}A_s - \frac{\boldsymbol{r}(\mathbf{j})\mathbf{R}(\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q}\overline{\alpha}_s\mathrm{d}A_s.
\end{aligned}
$$

Now we consider some special cases where the asymptotic (co)variances are simpler. As a consequence, the asymptotic variance estimators are also much simplified.

First, we consider the scenario $\overline{\alpha}_t \equiv 0$. The observations schemes that satisfy this condition include the *regular sampling scheme*, the *time-changed regular sampling scheme*; next, let $\overline{\alpha}_t \equiv 1$, one can verify that the *modulated Poisson sampling scheme* satisfies this condition, see the discussion in Jacod et al. (2017). The asymptotic (co)variance becomes

$$
\overline{\sigma}(\mathbf{j},\mathbf{j}')_t = \begin{cases} \frac{\boldsymbol{s}(\mathbf{j},\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q+q'}\mathrm{d}A_s, & \text{if } \overline{\alpha}_t \equiv 0;\\ \frac{\boldsymbol{s}(\mathbf{j},\mathbf{j}')+\boldsymbol{r}(\mathbf{j})\boldsymbol{r}(\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q+q'}\mathrm{d}A_s - \mathbf{R}(\mathbf{j})_t\mathbf{R}(\mathbf{j}')_t, & \text{if } \overline{\alpha}_t \equiv 1. \end{cases}
$$

and a consistent estimator is given by

$$
\widehat{\sigma}(\mathbf{j},\mathbf{j}')_t = \begin{cases} \frac{1}{n_t}\widehat{\sigma}_1(\mathbf{j},\mathbf{j}')_t^n, & \text{if } \overline{\alpha}_t \equiv 0;\\ \frac{1}{n_t}\left(\widehat{\sigma}_1(\mathbf{j},\mathbf{j}')_t^n + \widehat{\sigma}_2'(\mathbf{j},\mathbf{j}')_t^n + \widehat{\sigma}_3'(\mathbf{j},\mathbf{j}')_t^n\right), & \text{if } \overline{\alpha}_t \equiv 1; \end{cases}
$$

where

$$
\begin{aligned}
\widehat{\sigma}_2'(\mathbf{j},\mathbf{j}')_t^n &= \sum_{i=0}^{n_t - w(3)_n} \Delta_{\mathbf{j}}(Y)_{i+w(3)_2^n}^n \Delta_{\mathbf{j}'}(Y)_{i+w(3)_3^n}^n,\\
\widehat{\sigma}_3'(\mathbf{j},\mathbf{j}')_t^n &= -\frac{1}{n_t}\mathrm{ReMeDI}(Y;\mathbf{j})_t^n\mathrm{ReMeDI}(Y;\mathbf{j}')_t^n.
\end{aligned}
$$

References

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.

Athey, S. and G. W. Imbens (2006): "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, 74, 431–497.

Bandi, F. M., D. Pirino, and R. Reno (2017): "Excess idle time," *Econometrica*, 85, 1793–1846.

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.

Biais, B., P. Hillion, and C. Spatt (1995): "An empirical analysis of the limit order book and the order flow in the Paris Bourse," *Journal of Finance*, 50, 1655–1689.

Bollerslev, T., J. Li, and Y. Xue (2018): "Volume, Volatility and Public News Announcements," *Review of Economic Studies*, 85, 2005—2041.

Chaker, S. (2017): "On high frequency estimation of the frictionless price: The use of observed liquidity variables," *Journal of Econometrics*, 201, 127–143.

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.

Christensen, K., R. C. Oomen, and M. Podolskij (2014): "Fact or friction: Jumps at ultra high frequency," *Journal of Financial Economics*, 114, 576–599.

Clinet, S. and Y. Potiron (2017): "Estimation for high-frequency data under parametric market microstructure noise," Tech. rep.

DA, R. AND D. XIU (2019): "When moving-average models meet high-frequency data: Uniform inference on volatility," Tech. rep.

DELBAEN, F. AND W. SCHACHERMAYER (1994): "A general version of the fundamental theorem of asset pricing," *Mathematische Annalen*, 300, 463–520.

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.

FOUCAULT, T., M. PAGANO, A. ROELL, AND A. RÖELL (2013): *Market Liquidity: Theory, Evidence, and Policy*, Oxford University Press.

FRENCH, K. R. AND R. ROLL (1986): "Stock return variances: The arrival of information and the reaction of traders," *Journal of Financial Economics*, 17, 5–26.

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

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

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

HENDERSHOTT, T. AND A. J. MENKVELD (2014): "Price pressures," *Journal of Financial Economics*, 114, 405–423.

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.

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

JACOD, J. AND A. N. SHIRYAEV (2003): *Limit Theorems for Stochastic Processes*, vol. 288, Springer-Verlag Berlin.

LEHMANN, B. N. (1990): "Fads, martingales, and market efficiency," *Quarterly Journal of Economics*, 105, 1–28.

LI, Y., S. XIE, AND X. ZHENG (2016): "Efficient estimation of integrated volatility incorporating trading information," *Journal of Econometrics*, 195, 33–50.

LI, Z. M., R. J. LAEVEN, AND M. H. VELLEKOOP (2020): "Dependent microstructure noise and integrated volatility estimation from high-frequency data," *Journal of Econometrics*, 215, 536–558.

LI, Z. M. AND O. LINTON (2019): "A ReMeDI for microstructure noise," https://ssrn.com/abstract=3423607.

——— (2020): "Supplementary material for 'A ReMeDI for microstructure noise'," https://papers.ssrn.com/abstract=3688787.

LINTON, O. AND S. MAHMOODZADEH (2018): "Implications of high-frequency trading for security markets," *Annual Review of Economics*, 10, 237–259.

O'HARA, M. (1995): *Market Microstructure Theory*, vol. 108, Blackwell Cambridge, MA.

PARK, A. AND H. SABOURIAN (2011): "Herding and contrarian behavior in financial markets," *Econometrica*, 79, 973–1026.

PARLOUR, C. A. (1998): "Price dynamics in limit order markets," *Review of Financial Studies*, 11, 789–816.

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

TODOROV, V. AND G. TAUCHEN (2011): "Volatility jumps," *Journal of Business & Economic Statistics*, 29, 356–371.

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

YATCHEW, A. (1997): "An elementary estimator of the partial linear model," *Economics Letters*, 57, 135–143.

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.

Supplementary Materials for “A ReMeDI for Microstructure Noise”

Technical Proofs

Z. Merrick Li* Oliver Linton†

September 8, 2020

A Assumptions and Notations

In the proofs, K will be a constant that may change from line to line. When it depends on some parameter *par*, we write K_{par} instead. But it never depends on n or any parameters that depend on n.

Let V be any Itô semimartingale on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ that has a Grigelionis representation as X in (2) with coefficients $b^V, \sigma^V, \vartheta^V$, which satisfy:

Assumption (K). *The processes $b^V, \sigma^V, \vartheta^V$ are bounded with $\vartheta^V(\omega, t, z) \leq J(z)$ for some bounded function J on E satisfying $\int J^2(z)\lambda(\mathrm{d}z) < \infty$.*

Then for any V satisfying Assumption (K) and any $r \geq 2$, we have for any finite $(\mathcal{F}_t)$-stopping times $S \leq T$,

$$\mathbb{E}\left(|V_T - V_S|^r \,|\mathcal{F}_S\right) \leq \mathbb{E}\left(T - S\,|\mathcal{F}_S\right). \tag{A.1}$$

We also have

$$|\mathbb{E}\left(V_T - V_S\,|\mathcal{F}_S\right)| \leq \mathbb{E}\left(T - S\,|\mathcal{F}_S\right). \tag{A.2}$$

Assumption (SHON). *We have Assumption (H), (N), (O) and further assume that the processes $X, \alpha, \overline{\alpha}$ and γ satisfy Assumption (K), and the process $1/\alpha$ is bounded.*

*Faculty of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom. Email: z.merrick.li@gmail.com

†Faculty of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom. Email: obl20@cam.ac.uk.

According to a "localization procedure", see Lemma A.3 of Jacod et al. (2017), we can always assume (SHON) below, which implies the following

$$\exists\rho \in (1/2+\eta, 1),\ \delta(n,i) \le K\delta_n^{\rho}, \quad A_t \le Kt, \quad \mathbb{P}(\Omega_t^n) \to 1 \text{ if } \Omega_t^n := \{\delta_n n_t \le 1 + Kt\}. \tag{A.3}$$

In the sequel, we will assume $\mathbf{j} = (j_1, \ldots, , j_q), \mathbf{j}' = (j_1', \ldots, , j_{q'}') \in \mathfrak{J}$, and assume without loss of generality that $q \le q'$. Now we introduce some notations that will be used throughout the proofs. In the sequel, k_n and $\boldsymbol{k}_n$ are an integer and a vector of integers that will be specified later.

$$\begin{aligned} &g(\mathbf{j}, \boldsymbol{k}_n)_i^n := \Delta_{\mathbf{j}}^{\boldsymbol{k}_n}(Y)_i^n - (\gamma_i^n)^q \Delta_{\mathbf{j}}^{\boldsymbol{k}_n}(\chi)_i^n, \quad u(\mathbf{j}, \boldsymbol{k}_n)_i^n := \Delta_{\mathbf{j}}^{\boldsymbol{k}_n}(\chi)_i^n - \boldsymbol{r}(\mathbf{j}; \boldsymbol{k}_n), \quad q_n := 2^{q-1}k_n; \\ &\quad q_n' := 2^{q'-1}k_n, \quad d_i^n := \alpha_i^n \delta(n, i+1) - \delta_n, \quad \theta(\mathbf{j}, \boldsymbol{k}_n)_i^n := \sqrt{\delta_n}(\gamma_i^n)^q u(\mathbf{j}, \boldsymbol{k}_n)_i^n; \end{aligned} \tag{A.4}$$

where

$$\boldsymbol{r}(\mathbf{j}; \boldsymbol{k}_n) := \mathbb{E}\Big(\Delta_{\mathbf{j}}^{\boldsymbol{k}_n}(\chi)_0^n\Big). \tag{A.5}$$

When $\boldsymbol{k}_n$ satisfies the conditions specified in (14), we write $g(\mathbf{j})_i^n$ and $u(\mathbf{j})_i^n$ instead of $g(\mathbf{j}, \boldsymbol{k}_n)_i^n$ and $u(\mathbf{j}, \boldsymbol{k}_n)_i^n$.

Following Jacod et al. (2017), we introduce the decomposition of the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge0}, \mathbb{P})$. The processes $X, \alpha, \overline{\alpha}, \gamma$ and the observation times t_i^n are defined on a space $(\Omega^{(0)}, \mathcal{F}^{(0)}, (\mathcal{F}_t)_{t\ge0}, \mathbb{P}^{(0)})$; $\{\chi_i\}_{i\in\mathbb{Z}}$ is defined on another space $(\Omega^{(1)}, \mathcal{G}, (\mathcal{G}_i)_{i\in\mathbb{Z}}, \mathbb{P}^{(1)})$ with $\mathcal{G}_i := \sigma(\chi_k : k \le i)$ and $\mathcal{G}^i := \sigma(\chi_k : k \ge i)$. Let $\Omega = \Omega^{(0)} \times \Omega^{(1)}, \mathcal{F} = \mathcal{F}^{(0)} \otimes \mathcal{G}, \mathbb{P} = \mathbb{P}^{(0)} \otimes \mathbb{P}^{(1)}$.

B Some Auxiliary Results

Lemma B.1. *Let ξ, ξ' be two variables in the probability space $(\Omega^{(1)}, \mathcal{G}, \mathbb{P}^{(1)})$ so that ξ is $\mathcal{G}_i$-measurable and ξ' is $\mathcal{G}^{i+\ell}$-measurable, where $\ell \in \mathbb{N}_+$. Assume ξ, ξ' have bounded second moments. Under Assumption (N) in Li and Linton (2020),we have*

$$\big|\mathbb{E}(\xi\xi') - \mathbb{E}(\xi)\mathbb{E}(\xi')\big| \le K\ell^{-v}. \tag{B.1}$$

Proof. By first conditioning on $\mathcal{G}_i$, plus an application of the Cauchy-Schwarz inequality (for the first inequality), the boundedness of the second moments, and an application of Lemma VIII 3.102 of Jacod and Shiryaev (2003) (for the second inequality) yield:

$$\left|\mathbb{E}\left((\xi-\mathbb{E}(\xi))\left(\xi'-\mathbb{E}(\xi')\right)\right)\right| \leq \sqrt{\mathbb{E}\left((\xi-\mathbb{E}(\xi))^2\right)\mathbb{E}\left(\left(\mathbb{E}\left(\xi'-\mathbb{E}(\xi')\,|\mathcal{G}_i\right)\right)^2\right)} \leq K\ell^{-v}.$$

□

In the remaining part of this subsection, we will assume $\boldsymbol{k}_n$ satisfies (12). We denote

$$\underline{k}_n := \inf_{l\geq 2}(k_{l+1,n}-k_{l,n}); \quad \bar{k}_n := k_{q,n}\vee(-k_{1,n}) = \sup\{|k_{l,n}| : 1\leq l\leq q\}.$$

Lemma B.2. *We have under Assumption A.4 that*

$$|\boldsymbol{r}(\mathbf{j};\boldsymbol{k}_n)-\boldsymbol{r}(\mathbf{j})| \leq \frac{K}{(|k_{1,n}|\wedge \underline{k}_n)^{v}}, \tag{B.2}$$

where $\boldsymbol{r}(\mathbf{j})$ is defined in (9).

Proof. Let $\mathcal{Q}_q$ be the collection of all *proper* subsets of $\{1,2,\ldots,q\}$:

$$\mathcal{Q}_q = \{Q : Q \subsetneq \{1,\ldots,q\}\}, \tag{B.3}$$

thus for any $Q\in\mathcal{Q}_q$, $Q^c\neq\emptyset$. Now we have for $Q\in\mathcal{Q}_q$,

$$\left|\mathbb{E}\left(\prod_{l\in Q}\chi_{j_l}\prod_{l\in Q^c}\chi_{j_l-k_{l,n}}\right)\right| = \begin{cases}\left|\mathbb{E}\left(\chi_{j_1-k_{1,n}}\prod_{l\in Q}\chi_{j_l}\prod_{l\in Q^c,l\neq 1}\chi_{j_l-k_{l,n}}\right)\right| & \text{if } 1\in Q^c,\\ \left|\mathbb{E}\left(\chi_{j_{\bar{l}}-k_{\bar{l},n}}\prod_{l\in Q}\chi_{j_l}\prod_{l\in Q^c,l\neq \bar{l}}\chi_{j_l-k_{l,n}}\right)\right| & \text{if } 1\notin Q^c,\end{cases} \tag{B.4}$$

where $\bar{l}=\max\{l : l\in Q^c\}$ if $1\notin Q^c$.

Apply Lemma B.1 with $\xi=\prod_{l\in Q}\chi_{j_l}\prod_{l\in Q^c,l\neq 1}\chi_{j_l-k_{l,n}}$, $\xi'=\chi_{j_1-k_{1,n}}$, $i=j_1$, $\ell=|k_{1,n}|$ if $1\in Q^c$, and $\xi=\chi_{j_{\bar{l}}-k_{\bar{l},n}}$, $\xi'=\prod_{l\in Q}\chi_{j_l}\prod_{l\in Q^c,l\neq\bar{l}}\chi_{j_l-k_{l,n}}$, $i=j_{\bar{l}}-k_{\bar{l},n}$, $\ell=\underline{k}_n$ if $1\notin Q^c$, we readily get $\left|\mathbb{E}\left(\prod_{l\in Q}\chi_{j_l}\prod_{l\in Q^c}\chi_{j_l-k_{l,n}}\right)\right| \leq C\left(|k_{1,n}|\wedge\underline{k}_n\right)^{-v}$. Now (B.2) is proved since

$$\boldsymbol{r}(\mathbf{j};\boldsymbol{k}_n)-\boldsymbol{r}(\mathbf{j}) = \sum_{Q\in\mathcal{Q}_q}(-1)^{|Q^c|}\mathbb{E}\left(\prod_{l\in Q}\chi_{j_l}\prod_{l\in Q^c}\chi_{j_l-k_{l,n}}\right).$$

□

Lemma B.3. *Assume $(\bar{k}_n\vee j_1)\delta_n^{\rho}\to 0$, let*

$$\mathrm{ReMeDI}'(\chi;\mathbf{j},\boldsymbol{k}_n)_t^n := \sum_{i=2^{q-1}}^{n_t-k_n-j_1}(\gamma_i^n)^q\Delta_{\mathbf{j}}^{k_n}(\chi)_i^n. \tag{B.5}$$

Then for any $r > 1$, there is some constant $K_{r,q} > 0$ such that

$$\mathbb{E}\left(\left|\text{ReMeDI}(Y; \mathbf{j}, \boldsymbol{k}_n)^n_t - \text{ReMeDI}'(\chi; \mathbf{j}, \boldsymbol{k}_n)^n_t\right| \mathbf{1}_{\{\Omega^n_t\}}\right) \leq K_{r,q}(\bar{k}_n \vee j_1)^{\frac{1}{r}} \delta_n^{\frac{\rho}{r}-1}. \tag{B.6}$$

Proof. Let

$$\zeta^n_{i,l} := X^n_{i+j_l} - X^n_{i+j_l-k_{l,n}} + \left(\gamma^n_{i+j_l} - \gamma^n_i\right)\chi_{i+j_l} - \left(\gamma^n_{i+j_l-k_{l,n}} - \gamma^n_i\right)\chi_{i+j_l-k_{l,n}};$$
$$\zeta'^n_{i,l} := \gamma^n_i\left(\chi_{i+j_l} - \chi_{i+j_l-k_{l,n}}\right).$$

Now it follows (recall $\mathcal{Q}_q$ defined in (B.3)) that

$$\Delta^{\boldsymbol{k}_n}_{\mathbf{j}}(Y)^n_i = \prod_{l=1}^{q}(\zeta^n_{i,l} + \zeta'^n_{i,l}), \quad (\gamma^n_i)^q \Delta^{\boldsymbol{k}_n}_{\mathbf{j}}(\chi)^n_i = \prod_{l=1}^{q}\zeta'^n_{i,l}, \quad g(\mathbf{j}, \boldsymbol{k}_n)^n_i = \sum_{Q\in\mathcal{Q}_q}\prod_{l\in Q}\zeta'^n_{i,l}\prod_{l\in Q^c}\zeta^n_{i,l}. \tag{B.7}$$

Apply (A.1) for X and γ, and the fact that χ has bounded moments, we have for any $k \geq 2$

$$\mathbb{E}\left(|\zeta^n_{i,l}|^k\right) \leq K(\bar{k}_n \vee j_1)\delta^\rho_n, \quad \mathbb{E}\left(|\zeta'^n_{i,l}|^k\right) \leq K, \quad \forall i, l. \tag{B.8}$$

Let $\ell = |Q^c|$, whence $\ell \geq 1$ (recall (B.3)). For $r \geq 2$, apply Hölder's inequality with exponents $(\underbrace{r\ell, \ldots, r\ell}_{\ell}, \frac{r}{r-1})$, we have

$$\mathbb{E}\left(\left|\prod_{l\in Q^c}\zeta^n_{i,l}\prod_{l\in Q}\zeta'^n_{i,l}\right|\right) \leq \prod_{l\in Q^c}\left(\mathbb{E}\left(|\zeta^n_{i,l}|^{r\ell}\right)\right)^{\frac{1}{r\ell}}\left(\mathbb{E}\left(\left|\prod_{l\in Q}\zeta'^n_{i,l}\right|^{\frac{r}{r-1}}\right)\right)^{\frac{r-1}{r}} \overset{\text{(B.8)}}{\leq} K_{r,q}\left((\bar{k}_n \vee j_1)\delta^\rho_n\right)^{\frac{1}{r}}. \tag{B.9}$$

(B.6) follows immediately.

For $1 < r \leq 2$, we first note (B.9) still holds if $\ell \geq 2$. For $\ell = 1$, we let $Q^c = \{l^*\}$. Let $\mathcal{H}^n_{i,l^*} := \mathcal{F}^n_{i+j_{l^*}-k_{l^*,n}} \otimes \mathcal{G}$ if $l^* > 1$, and $\mathcal{H}^n_{i,l^*} := \mathcal{F}^n_i \otimes \mathcal{G}$ if $l^* = 1$. Then by the independence of $\mathcal{G}$ and $\mathcal{F}^{(0)}$, (A.2) for γ , we have

$$\left|\mathbb{E}\left(\zeta^n_{i,l^*}\,|\mathcal{H}^n_{i,l^*}\right)\right| \leq K(\bar{k}_n \vee j_1)\delta^\rho_n\left(1 + |\chi_{i+j_{l^*}}| + \left|\chi_{i+j_{l^*}-k_{l^*,n}}\right|\right). \tag{B.10}$$

Since $\prod_{l\neq l^*}\zeta'^n_{i,l}$ is measurable with respect to $\mathcal{H}^n_{i,l^*}$, the boundedness of γ, the fact that

χ has bounded moments of all orders and (B.10) imply

$$\begin{aligned}
&\mathbb{E}\left(\left|\mathbb{E}\left(\zeta^n_{i,l^*}\prod_{l\neq l^*}\zeta'^n_{i,l}\,|\mathcal{H}^n_{i,l^*}\right)\right|\right)\\
\leq &K(\bar{k}_n\vee j_1)\delta_n^{\rho}\mathbb{E}\left(\left(1+\left|\chi_{i+j_{l^*}}\right|+\left|\chi_{i+j_{l^*}-k_{l^*,n}}\right|\right)\prod_{l\neq l^*}\left|\chi_{i+j_l}-\chi_{i+j_l-k_{l,n}}\right|\right)\\
\leq &K(\bar{k}_n\vee j_1)\delta_n^{\rho}.
\end{aligned}\tag{B.11}$$

On the other hand, since $r>1$, apply Hölder's inequality, we get

$$\mathbb{E}\left(\left(\zeta^n_{i,l^*}\prod_{l\neq l^*}\zeta'^n_{i,l}\right)^2\right)\leq K\left((\bar{k}_n\vee j_1)\delta_n^{\rho}\right)^{1/r}. \tag{B.12}$$

Note that $\zeta^n_{i,l^*}\prod_{l\neq l^*}\zeta'^n_{i,l}$ is measurable with respect to $\mathcal{F}^n_{i+j_1-k_{1,n}}\otimes\mathcal{G}$, combined with (B.11) and (B.12), we can apply Lemma A.6 of Jacod et al. (2017) and obtain

$$\mathbb{E}\left(\left|\sum_{i=2^{q}-1}^{n_t-k_n-j_1}\zeta^n_{i,l^*}\prod_{l\neq l^*}\zeta'^n_{i,l}\right|\right)\leq K\left((\bar{k}_n\vee j_1)\delta_n^{\rho-1}+(\bar{k}_n\vee j_1)^{\frac{r+1}{2r}}\delta_n^{\frac{\rho-r}{2r}}\right)\leq K(\bar{k}_n\vee j_1)^{\frac{1}{r}}\delta_n^{\frac{\rho}{r}-1}.$$

The last inequality follows from the fact that $\delta_n^{\rho}(\bar{k}_n\vee j_1)\to 0$. Now the proof is complete. □

C Additional Auxiliary Results

In this section, we assume $\boldsymbol{k}_n$ is specified as follows for given an integer k_n: $k_{l,n}=-k_n$ if $l=1$, $k_{l,n}=2^{l-1}k_n$ if $l\geq 2$. In line with the notations in Li and Linton (2020), we will write $\Delta_{\mathbf{j}}(\cdot)^n_i$ instead of $\Delta^{\boldsymbol{k}_n}_{\mathbf{j}}(\cdot)^n_i$ when $\boldsymbol{k}_n$ is specified as above. Moreover, we will replace $\boldsymbol{r}(\mathbf{j},\boldsymbol{k}_n)$ (recall (A.5)) by $\boldsymbol{r}(\mathbf{j},k_n)$. We further denote $h_{\mathbf{j}}(i,l)_n:=i+j_l-k_{l,n}$. For $Q_q\subset\{1,2,\ldots,q\}$, let

$$\widetilde{\chi}(Q_q,\mathbf{j})^n_i:=\begin{cases}\prod_{j_l\in\mathbf{j}}\chi_{i+j_l}-\boldsymbol{r}(\mathbf{j}) & \text{if } Q_q^c=\emptyset;\\ (-1)^{|Q_q^c|}\prod_{l\in Q_q}\chi_{i+j_l}\prod_{l\in Q_q^c}\chi_{h_{\mathbf{j}}(i,l)_n} & \text{if } Q_q^c\neq\emptyset.\end{cases}$$

$\widetilde{\chi}(Q'_{q'}, \mathbf{j}')^n_i$ is defined in a similar manner for $\mathbf{j}', Q'_{q'} \subset \{1, 2, \ldots, q'\}$. We have for any i, k that (recall $u(\mathbf{j})^n_i$ defined in (A.4)):

$$u(\mathbf{j})^n_i = \sum_{Q_q \subset \{1,2,\ldots,q\}} \widetilde{\chi}(Q_q, \mathbf{j})^n_i; \quad u(\mathbf{j}')^n_{i+k} = \sum_{Q'_{q'} \subset \{1,2,\ldots,q'\}} \widetilde{\chi}(Q'_{q'}, \mathbf{j}')^n_{i+k}.$$

Now we introduce four mutually exclusive categories of pairs of $(Q_q, Q'_{q'})$, or their complements $(Q^c_q, (Q'_{q'})^c)$:

$$\begin{cases} Q^c_q = (Q'_{q'})^c, & \text{(C.1)} \\ Q^c_q = \{l\}, (Q'_{q'})^c = \{l'\}, l \neq l', & \text{(C.2)} \\ (Q'_{q'})^c = \emptyset, Q^c_q = \{l\}, & \text{(C.3)} \\ Q^c_q = \emptyset, (Q'_{q'})^c = \{l'\}. & \text{(C.4)} \end{cases}$$

First, we show

Lemma C.1. *For any pair $(Q_q, Q'_{q'})$ that does* not *satisfy (C.1) to (C.4) , we define the following sets of indices for any integers i, k:*

$$\mathbb{I}(Q^c_q)_i = \{h_{\mathbf{j}}(i, l)_n : l \in Q^c_q\}, \quad \mathbb{I}((Q'_{q'})^c)_{i+k} = \{h_{\mathbf{j}'}(i+k, l')_n : l' \in (Q'_{q'})^c\};$$
$$\mathbb{I}(Q_q)_i = \{i + j_l : l \in Q_q\}, \qquad \mathbb{I}(Q'_{q'})_{i+k} = \{i + k + j'_{l'} : l' \in Q'_{q'}\}.$$

Then there exists at least one index in $\mathbb{I}(Q^c_q)_i \cup \mathbb{I}((Q'_{q'})^c)_{i+k}$ that is at least $k_n/3$ apart from the remaining indices in $\mathbb{I}(Q^c_q)_i \cup \mathbb{I}((Q'_{q'})^c)_{i+k} \cup \mathbb{I}(Q_q)_i \cup \mathbb{I}(Q'_{q'})_{i+k}$.

Proof. We first consider pairs of $(Q_q, Q'_{q'})$ that do *not* satisfy (C.1) to (C.4) but *satisfy* $|Q^c_q| = |(Q'_{q'})^c|$. If this were true, then violating (C.1) and (C.2) implies $|Q^c_q| = |(Q'_{q'})^c| \geq 2$. Now suppose Lemma C.1 is not true. Denote $(l_\tau)_{1 \leq \tau \leq |Q^c_q|}$ so that $h_{\mathbf{j}}(i, l_\tau)_n \in \mathbb{I}(Q^c_q)_i$, and they are in an ascending order, i.e., $h_{\mathbf{j}}(i, l_1)_n < h_{\mathbf{j}}(i, l_2)_n < \cdots < h_{\mathbf{j}}(i, l_{|Q^c_q|})_n$, or equivalently, $l_1 > l_2 > \cdots > l_{|Q^c_q|}$.[1] $(l'_{\tau'})_{1 \leq \tau' \leq |(Q'_{q'})^c|}$ are defined similarly for the indices $h_{\mathbf{j}'}(i + k, l'_{\tau'})_n \in \mathbb{I}((Q'_{q'})^c)_{i+k}$. Since the minimal distance between any index in $\mathbb{I}(Q^c_q)_i$ (or $\mathbb{I}((Q'_{q'})^c)_{i+k}$) and the remaining indices in $\mathbb{I}(Q^c_q)_i \cup \mathbb{I}(Q_q)_i$ (or $\mathbb{I}((Q'_{q'})^c)_{i+k} \cup \mathbb{I}((Q'_{q'}))_{i+k}$) is k_n, we conclude that each pair of indices

$$\left| h_{\mathbf{j}}(i, l_\tau)_n - h_{\mathbf{j}'}(i + k, l'_\tau)_n \right| \leq k_n/3, \tau = 1, \ldots, |Q^c_q|, \tag{C.5}$$

were Lemma C.1 not true.

[1]In this proof, many inequalities hold up to adding a constant. For example, we conclude $z_n > z'_n$ if $z_n + c_1 > z'_n + c_2$, where c_1, c_2 are some constant and z_n, z'_n are large when n is large.

Assume $l_1 > l'_1$. Then by (C.5) we have (since $|Q_q^c| \geq 2$) :

$$\left|\left(h_{\mathbf{j}}(i,l_1)_n - h_{\mathbf{j}}(i,l_2)_n\right) - \left(h_{\mathbf{j}'}(i+k,l'_1)_n - h_{\mathbf{j}'}(i+k,l'_2)_n\right)\right| \leq 2k_n/3, \tag{C.6}$$

which implies

$$\begin{cases} 2^{l_1-1} = 2^{l'_1-1}, & \text{if } l_2 = 1, l'_2 = 1; \\ 2^{l_1-1} = 2^{l'_1-1} - 2^{l'_2-1} - 1, & \text{if } l_2 = 1, l'_2 > 1; \\ 2^{l_1-1} = 2^{l_2-1} + 2^{l'_1-1} + 1, & \text{if } l_2 > 1, l'_2 = 1; \\ 2^{l_1-1} = 2^{l_2-1} + 2^{l'_1-1} - 2^{l'_2-1}, & \text{if } l_2 > 1, l'_2 > 1. \end{cases}$$

But it contradicts the fact that $l_1 > \max(l_2, l'_1, l'_2)$. Therefore we have $l_1 \leq l'_1$; similarly we get $l_1 \geq l'_1$, thus we conclude $l_1 = l'_1$. We also have $l_2 = l'_2$ since (C.5) (with $\tau = 1$) implies $|k| \leq k_n/3$. We can proceed to prove $l_\tau = l'_\tau$ for all $l_\tau \in Q_q^c$, i.e., $Q_q^c = (Q'_{q'})^c$, which is a contradiction. Therefore, we conclude that for any pair of $(Q_q, Q'_{q'})$ that does *not* satisfy (C.1) to (C.4), we have $|Q_q^c| \neq |(Q'_{q'})^c|$.

Now we consider pairs of $(Q_q, Q'_{q'})$ that do *not* satisfy (C.1) to (C.4) but satisfy $|Q_q^c| > |(Q'_{q'})^c|$. (C.3) implies $|Q_q^c| \geq 2$. Consider the following scenarios:

1. If $|Q_q^c| > |(Q'_{q'})^c| + 1$, apply the *Pigeonhole Principle*: consider $|Q_q^c|$ "containers" centred at $\{h_{\mathbf{j}}(i,l)_n : h_{\mathbf{j}}(i,l)_n \in \mathbb{I}(Q_q^c)_i\}$ with "radius" $k_n/3$. Were Lemma C.1 not true, we need to place the $|(Q'_{q'})^c| + 1$ "items" $\{h_{\mathbf{j}'}(i+k,l')_n \in \mathbb{I}((Q'_{q'})^c)_{i+k}, \mathbb{I}(Q'_{q'})_{i+k}\}$[2] into the "containers". The Pigeonhole Principle implies at least one of the "containers" is empty, thus Lemma C.1 must be true.

2. If $|Q_q^c| = |(Q'_{q'})^c| + 1 \geq 2$ and Lemma C.1 is false, there is one-to-one correspondence between the $|(Q'_{q'})^c| + 1$ "items" $\{h_{\mathbf{j}'}(i+k,l')_n \in \mathbb{I}((Q'_{q'})^c)_{i+k}, \mathbb{I}(Q'_{q'})_{i+k}\}$ and $|Q_q^c|$ "items" $\{h_{\mathbf{j}}(i,l)_n : l \in Q_q^c\}$ so that each pair has a distance less than $k_n/3$ (recall a representation of such correspondence by (C.5)). Now we need to consider the following two cases.

   (a) $\mathbb{I}((Q'_{q'})^c)_{i+k} = \{h_{\mathbf{j}}(i+k,1)_n\}$, i.e., $(Q'_{q'})^c = \{1\}$. Let's fix the index of $\mathbb{I}(Q'_{q'})_{i+k}$ at $i+k$.[3] Let $Q_q^c = \{l_1, l_2\}$. Similar to (C.6), we have

$$\left|i + k + j'_1 + k_n - (i+k) - (h_{\mathbf{j}}(i,l_1)_n - h_{\mathbf{j}}(i,l_2)_n)\right| \leq 2k_n/3.$$

[2] Asymptotically we treat $\mathbb{I}(Q'_{q'})_{i+k}$ as one "item" since the distances between the indices in $\mathbb{I}(Q'_{q'})_{i+k}$ are independent of n, thus "fixed".

[3] It can be any of $\{i+k+j'_{l'} : l' \in Q'_{q'}\}$, but asymptotically they are equivalent.

This contradicts to

$$\left|i+k+j_1'+k_n-(i+k)-(h_{\mathbf{j}}(i,l_1)-h_{\mathbf{j}}(i,l_2))\right| \geq \left|(h_{\mathbf{j}}(i,l_1)_n-h_{\mathbf{j}}(i,l_2)_n)\right|-k_n,$$

which is no smaller than k_n

(b) Now assume $\exists l' \in (Q'_{q'})^c, l' > 1$, we can apply the arguments used above to show for each $l'_\tau > 1, l'_\tau \in (Q'_{q'})^c$, there is some $l_\tau \in Q_q^c$ such that $l'_\tau = l_\tau$. We also conclude $|k| \leq k_n/3$. Let l^*, l^*_+ be the two indices satisfying (recall $|Q_q^c| \geq 2$) :

$$l^* = \operatorname*{argmax}_{\{l>1: l\in (Q'_{q'})^c\}} h_{\mathbf{j}}(i,l)_n; \quad l^*_+ = \operatorname*{argmax}_{\{l\neq l^*: h_{\mathbf{j}}(i,l)_n \in \mathbb{I}(Q_q^c)_i\}} h_{\mathbf{j}}(i,l)_n.$$

(Note that l^*_+ could be 1.) Now we have

$$\left|h_{\mathbf{j}'}(i+k,l^*)_n-(i+k)-(h_{\mathbf{j}}(i,l^*)_n-h_{\mathbf{j}}(i,l^*_+)_n)\right| \leq 2k_n/3.$$

But this contradicts to

$$\begin{aligned} &\left|h_{\mathbf{j}'}(i+k,l^*)_n-(i+k)-(h_{\mathbf{j}}(i,l^*)_n-h_{\mathbf{j}}(i,l^*_+)_n)\right| \\ \geq &\begin{cases} \left|(h_{\mathbf{j}}(i,l^*)_n-h_{\mathbf{j}}(i,l^*_+)_n)\right|-\left|h_{\mathbf{j}'}(i+k,l^*)_n-(i+k)\right| & \text{if } l^*_+=1 \\ \left|h_{\mathbf{j}'}(i+k,l^*)_n-(i+k)\right|-\left|(h_{\mathbf{j}}(i,l^*)_n-h_{\mathbf{j}}(i,l^*_+)_n)\right| & \text{if } l^*_+>1 \end{cases} \\ \geq &k_n. \end{aligned}$$

This finishes the proof of Lemma C.1 for the case $|Q_q^c| > |(Q'_{q'})^c|$. The conclusion for $|Q_q^c| < |(Q'_{q'})^c|$ can be proved analogously, and the proof now is complete. □

Lemma C.2. *For any pair $(Q_q, Q'_{q'})$ that does* not *satisfy* (C.1) *to* (C.4)*, we have*

$$\left|\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})_i^n \widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n\right)\right| \leq C k_n^{-v}, \quad \forall k \in \mathbb{Z}. \tag{C.7}$$

Proof. Let one of the indices satisfying Lemma C.1 be h^*. Write

$$\widetilde{\chi}(Q_q,\mathbf{j})_i^n \widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n = \hat{\chi}^n_{h^*-\frac{k_n}{3}} \chi_{h^*} \hat{\chi}^n_{h^*+\frac{k_n}{3}},$$

where $\hat{\chi}^n_{h^*-\frac{k_n}{3}}$ and $\hat{\chi}^n_{h^*+\frac{k_n}{3}}$ are the products of the remaining factors in $\widetilde{\chi}(Q_q,\mathbf{j})_i^n \widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n$ (other than χ_{h^*}) that are measurable with respect to $\mathcal{G}_{h^*-[k_n/3]}$ and $\mathcal{G}^{h^*+[k_n/3]}$, respectively. Since $\hat{\chi}^n_{h^*-\frac{k_n}{3}}, \chi_{h^*}$ and $\hat{\chi}^n_{h^*+\frac{k_n}{3}}$ are integrable, we can apply Lemma B.1 to get

$$\left|\mathbb{E}\left(\hat{\chi}^n_{h^*-\frac{k_n}{3}}\chi_{h^*}\hat{\chi}^n_{h^*+\frac{k_n}{3}}\right) - \mathbb{E}\left(\hat{\chi}^n_{h^*-\frac{k_n}{3}}\chi_{h^*}\right)\mathbb{E}\left(\hat{\chi}^n_{h^*+\frac{k_n}{3}}\right)\right| \leq Kk_n^{-v}; \text{ and } \left|\mathbb{E}\left(\hat{\chi}^n_{h^*-\frac{k_n}{3}}\chi_{h^*}\right)\right| \leq Kk_n^{-v}. \tag{C.8}$$

This finishes the proof of (C.7). □

Lemma C.3. *For all pairs of $(Q_q, Q'_{q'})$ that* satisfy (C.1)*, we have for any $k \in \mathbb{Z}$,*

$$\left|\sum_{(Q_q,Q'_{q'})}^{(C.1)} \mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})^n_i\widetilde{\chi}(Q'_{q'},\mathbf{j}')^n_{i+k}\right) - \boldsymbol{s}_0(\mathbf{j},\mathbf{j}';k) - \boldsymbol{s}_1(\mathbf{j},\mathbf{j}';k)\right| \leq Kk_n^{-v}. \tag{C.9}$$

Proof. If $Q_q^c = (Q'_{q'})^c = \emptyset$, we have $\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})^n_i\widetilde{\chi}(Q'_{q'},\mathbf{j}')^n_{i+k}\right) = \boldsymbol{s}_0(\mathbf{j},\mathbf{j}';k)$.

Now consider $Q_q^c = (Q'_{q'})^c \neq \emptyset$ so that $Q'_{q'} = Q_{q'}$ (recall $Q_{q'}$ defined in (B.5)), and

$$\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})^n_i\widetilde{\chi}(Q'_{q'},\mathbf{j}')^n_{i+k}\right) = \mathbb{E}\left(\prod_{l\in Q_q}\chi_{i+j_l}\prod_{l'\in Q_{q'}}\chi_{i+k+j'_{l'}}\prod_{l\in Q_q^c}\chi_{h_{\mathbf{j}}(i,l)_n}\chi_{h_{\mathbf{j}'}(i+k,l)_n}\right).$$

Let $|k| \leq \frac{k_n}{2}$, by successive conditioning as we did to obtain (C.8), we obtain

$$\left|\mathbb{E}\left(\prod_{l\in Q_q^c}\chi_{h_{\mathbf{j}}(i,l)_n}\chi_{h_{\mathbf{j}'}(i+k,l)_n}\right) - \prod_{l\in Q_q^c}\boldsymbol{r}(j_l,j'_l+k)\right| \leq Kk_n^{-v}.$$

This yields, together with the fact that $\left|\boldsymbol{r}\left(\mathbf{j}_{Q_q}\oplus\mathbf{j}'_{Q_{q'}}(+k)\right)\right|$ is bounded that

$$\begin{aligned}&\left|\mathbb{E}\left(\prod_{l\in Q_q}\chi_{i+j_l}\prod_{l'\in Q_{q'}}\chi_{i+k+j'_{l'}}\prod_{l\in Q_q^c}\chi_{h_{\mathbf{j}}(i,l)_n}\chi_{h_{\mathbf{j}'}(i+k,l)_n}\right) - \boldsymbol{r}\left(\mathbf{j}_{Q_q}\oplus\mathbf{j}'_{Q_{q'}}(+k)\right)\prod_{l\in Q_q^c}\boldsymbol{r}(j_l,j'_l+k)\right|\\ \leq &\left|\mathbb{E}\left(\left(\prod_{l\in Q_q}\chi_{i+j_l}\prod_{l'\in Q_{q'}}\chi_{i+k+j'_{l'}} - \boldsymbol{r}\left(\mathbf{j}_{Q_q}\oplus\mathbf{j}'_{Q_{q'}}(+k)\right)\right)\prod_{l\in Q_q^c}\chi_{h_{\mathbf{j}}(i,l)_n}\chi_{h_{\mathbf{j}'}(i+k,l)_n}\right)\right| + Kk_n^{-v}.\end{aligned}$$

Apply Lemma B.1, by successive conditioning, we get that the expectation after the inequality is also bounded by Kk_n^{-v}, since the indices $\{i+j_l, i+k+j'_{l'} : l \in Q_q, l' \in Q'_{q'}\}$ are at least $k_n/2$ apart from the indices $\{h_{\mathbf{j}}(i,l)_n, h_{\mathbf{j}'}(i+k,l)_n : l \in Q_q^c\}$. This proves

$$\left|\sum_{Q_q^c=(Q'_{q'})^c\neq\emptyset}\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})^n_i\widetilde{\chi}(Q'_{q'},\mathbf{j}')^n_{i+k}\right) - \boldsymbol{s}_1(\mathbf{j},\mathbf{j}';k)\right| \leq Kk_n^{-v}, \tag{C.10}$$

for $|k| \le k_n/2$. For $|k| \ge k_n/2$, we also have $\left|\mathbb{E}\left(\widetilde{\chi}(Q_q, \mathbf{j})_i^n \widetilde{\chi}(Q'_{q'}, \mathbf{j}')_{i+k}^n\right)\right| \le K k_n^{-v}$ and $|s_1(\mathbf{j}, \mathbf{j}'; k)| \le K k_n^{-v}$, thus (C.10) holds for $|k| \ge k_n/2$ as well. This completes the proof. □

Lemma C.4. *For all pairs $(Q_q, Q'_{q'})$ that satisfy (C.2) to (C.4), we have*

$$\left| \sum_{(Q_q, Q'_{q'})}^{(C.2)\sim(C.4)} \mathbb{E}\left(\widetilde{\chi}(Q_q, \mathbf{j})_i^n \widetilde{\chi}(Q'_{q'}, \mathbf{j}')_{i+k}^n\right) - s_{2,k_n}(\mathbf{j}, \mathbf{j}'; k) \right| \le K k_n^{-v}, \tag{C.11}$$

where

$$\begin{aligned} s_{2,k_n}(\mathbf{j}, \mathbf{j}'; k) := & \sum_{\substack{j_l \in \mathbf{j}, j'_{l'} \in \mathbf{j}' \\ l \neq l'}} \boldsymbol{r}_{k_n}(j_l, j'_{l'} + k) \boldsymbol{r}(\mathbf{j}_{-l}) \boldsymbol{r}(\mathbf{j}'_{-l'}) - \sum_{j_l \in \mathbf{j}} \boldsymbol{r}_{k_n}(\{j_l\} \oplus \mathbf{j}'(+k)) \boldsymbol{r}(\mathbf{j}_{-l}) \\ & - \sum_{j'_{l'} \in \mathbf{j}'} \boldsymbol{r}_{k_n}(\{j'_{l'} + k\} \oplus \mathbf{j}) \boldsymbol{r}(\mathbf{j}'_{-l'}), \end{aligned}$$

with

$$\boldsymbol{r}_{k_n}(j_l, j'_{l'} + k) := \begin{cases} \boldsymbol{r}(j_l, j'_{l'} + k - (2^{l'-1} + 1)k_n) & \text{if } l = 1, l' > 1 \\ \boldsymbol{r}(j_l, j'_{l'} + k + (2^{l-1} + 1)k_n) & \text{if } l' = 1, l > 1 \\ \boldsymbol{r}(j_l, j'_{l'} + k - (2^{l'-1} - 2^{l-1})k_n) & \text{if } l > 1, l' > 1, l \neq l' \end{cases},$$

$$\boldsymbol{r}_{k_n}(\{j_l\} \oplus \mathbf{j}'(+k)) := \begin{cases} \boldsymbol{r}(\{j_l\} \oplus \mathbf{j}'(+(k - k_n))) & \text{if } l = 1 \\ \boldsymbol{r}(\{j_l\} \oplus \mathbf{j}'(+(k + 2^{l-1}k_n))) & \text{if } l > 1 \end{cases},$$

$$\boldsymbol{r}_{k_n}(\{j'_{l'} + k\} \oplus \mathbf{j}) := \begin{cases} \boldsymbol{r}(\{j'_{l'} + k\} \oplus \mathbf{j}(+(-k_n))) & \text{if } l' = 1 \\ \boldsymbol{r}(\{j'_{l'} + k\} \oplus \mathbf{j}(+(2^{l-1}k_n))) & \text{if } l' > 1 \end{cases}.$$

Proof. First, we prove

$$\left| \sum_{(Q_q, Q'_{q'})}^{(C.2)} \mathbb{E}\left(\widetilde{\chi}(Q_q, \mathbf{j})_i^n \widetilde{\chi}(Q'_{q'}, \mathbf{j}')_{i+k}^n\right) - \sum_{j_l \in \mathbf{j}, j'_{l'} \in \mathbf{j}', l \neq l'} \boldsymbol{r}_{k_n}(j_l, j'_{l'} + k) \boldsymbol{r}(\mathbf{j}_{-l}) \boldsymbol{r}(\mathbf{j}'_{-l'}) \right| \le K k_n^{-v}. \tag{C.12}$$

Let's assume $l' > l = 1$. Then for $\left(\frac{1}{2} + 2^{l'-1}\right) k_n \le k \le \left(\frac{3}{2} + 2^{l'-1}\right) k_n$, $\chi_{h_{\mathbf{j}}(i,l)_n} \chi_{h_{\mathbf{j}'}(i+k,l')}$, $\prod_{\bar{l} \neq p} \chi_{i+j_{\bar{l}}}$ and $\prod_{\bar{l}' \neq l'} \chi_{i+k+j'_{\bar{l}'}}$ are asymptotically at least $k_n/2$ away from each other. Apply Lemma B.1, we can separate the terms with an error bounded by Kk_n^{-v}:

$$\left|\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})_i^n\widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n\right)-\boldsymbol{r}(j_l,j'_{l'}+k-(2^{l'-1}+1)k_n)\boldsymbol{r}(\mathbf{j}_{-l})\boldsymbol{r}(\mathbf{j}'_{-l'})\right|\leq Kk_n^{-v}. \quad \text{(C.13)}$$

For $k<\left(\frac{1}{2}+2^{l'-1}\right)k_n$ or $k>\left(\frac{3}{2}+2^{l'-1}\right)k_n$, at least one of $h_{\mathbf{j}}(i,l)_n,h_{\mathbf{j}'}(i+k,l')$ is at least $k_n/2$ from the remaining factors in $\widetilde{\chi}(Q_q,\mathbf{j})_i^n\widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n$, thus we can show

$$\left|\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})_i^n\widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n\right)\right|\vee\left|\boldsymbol{r}(j_l,j'_{l'}+k-(2^{l'-1}+1)k_n)\right|\leq Kk_n^{-v},$$

thus (C.13) still holds. Similarly, we have for $l>l'=1$ that

$$\left|\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})_i^n\widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n\right)-\boldsymbol{r}(j_l,j'_{l'}+k+(2^{l-1}+1)k_n)\boldsymbol{r}(\mathbf{j}_{-l})\boldsymbol{r}(\mathbf{j}'_{-l'})\right|\leq Kk_n^{-v}, \quad \text{(C.14)}$$

for $-\left(\frac{3}{2}+2^{l-1}\right)k_n\leq k\leq-\left(\frac{1}{2}+2^{l-1}\right)k_n$, and

$$\left|\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})_i^n\widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n\right)\right|\vee\left|\boldsymbol{r}(j_l,j'_{l'}+k+(2^{l-1}+1)k_n)\right|\leq Kk_n^{-v},$$

for $k<-\left(\frac{3}{2}+2^{l-1}\right)k_n$ or $k>-\left(\frac{1}{2}+2^{l-1}\right)k_n$. Now assume $l'\neq l, l>1, l'>1$. For $\left(2^{l'-1}-2^{l-1}-\frac{1}{2}\right)k_n\leq k\leq\left(2^{l'-1}-2^{l-1}+\frac{1}{2}\right)k_n$, we have

$$\left|\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})_i^n\widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n\right)-\boldsymbol{r}(j_l,j'_{l'}+k-(2^{l'-1}-2^{l-1})k_n)\boldsymbol{r}(\mathbf{j}_{-l})\boldsymbol{r}(\mathbf{j}'_{-l'})\right|\leq Kk_n^{-v}.$$

For $k>\left(2^{l'-1}-2^{l-1}+\frac{1}{2}\right)k_n$ or $k<\left(2^{l'-1}-2^{l-1}-\frac{1}{2}\right)k_n$, we have

$$\left|\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})_i^n\widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n\right)\right|\vee\left|\boldsymbol{r}(j_l,j'_{l'}+k-(2^{l'-1}-2^{l-1})k_n)\right|\leq Kk_n^{-v}.$$

This completes the proof of (C.12).

The proofs of

$$\left|\sum_{(Q_q,Q'_{q'})}^{\text{(C.3)}}\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})_i^n\widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n\right)+\sum_{j_l\in\mathbf{j}}\boldsymbol{r}_{k_n}(\{j_l\}\oplus\mathbf{j}'(+k))\boldsymbol{r}(\mathbf{j}_{-l})\right|\leq Kk_n^{-v},$$

$$\left|\sum_{(Q_q,Q'_{q'})}^{\text{(C.4)}}\mathbb{E}\left(\widetilde{\chi}(Q_q,\mathbf{j})_i^n\widetilde{\chi}(Q'_{q'},\mathbf{j}')_{i+k}^n\right)+\sum_{j'_{l'}\in\mathbf{j}'}\boldsymbol{r}_{k_n}(\{j'_{l'}+k\}\oplus\mathbf{j})\boldsymbol{r}(\mathbf{j}'_{-l'})\right|\leq Kk_n^{-v},$$

are similar (in fact, simpler), and this completes the proof. □

Lemma C.5. *For any integers i, k, we have*

$$\left|\mathbb{E}\left(u(\mathbf{j})_i^n u(\mathbf{j}')_{i+k}^n\right) - s_{k_n}\left(\mathbf{j}, \mathbf{j}'; k\right)\right| \leq K k_n^{-v}, \tag{C.15}$$

where

$$s_{k_n}\left(\mathbf{j}, \mathbf{j}'; k\right) := s_0\left(\mathbf{j}, \mathbf{j}'; k\right) + s_1\left(\mathbf{j}, \mathbf{j}'; k\right) + s_{2,k_n}\left(\mathbf{j}, \mathbf{j}'; k\right),$$

and $s_0\left(\mathbf{j}, \mathbf{j}'; k\right), s_1\left(\mathbf{j}, \mathbf{j}'; k\right)$ are introduced in Appendix B.1 in Li and Linton (2020).

Proof. (B.2) implies we can replace $\boldsymbol{r}(\mathbf{j}; k_n), \boldsymbol{r}(\mathbf{j}'; k_n)$ by $\boldsymbol{r}(\mathbf{j}), \boldsymbol{r}(\mathbf{j}')$ with errors no larger than $K k_n^{-v}$. Now (C.15) follows from (C.9) and (C.11). □

Next, we will present and prove a key result on stable convergence.

Theorem C.1. *Let*

$$G_t^n := \sum_{i=q_n}^{n_t - k_n - j_1} \theta(\mathbf{j}, \boldsymbol{k}_n), \quad G_t'^n := \sum_{i=q_n'}^{n_t - k_n - j_1'} \theta(\mathbf{j}', \boldsymbol{k}_n);$$
$$H_t^n := \frac{1}{\sqrt{\delta_n}} \sum_{i=q_n}^{n_t - k_n - j_1} (\gamma_i^n)^q d_i^n, \quad H_t'^n := \frac{1}{\sqrt{\delta_n}} \sum_{i=q_n'}^{n_t - k_n - j_1'} (\gamma_i^n)^{q'} d_i^n;$$
$$\mathbf{G}_t^n := (G_t^n, G_t'^n), \quad \mathbf{H}_t^n := (H_t^n, H_t'^n).$$

Assume (14), we have $(\mathbf{G}_t^n, \mathbf{H}_t^n)$ converges $\mathcal{F}_\infty$-stably in law to $(\mathbf{G}_t, \mathbf{H}_t)$ with components $\mathbf{G}_t = (G_t, G_t'), \mathbf{H}_t = (H_t, H_t')$ that is defined on an extension $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ of $(\Omega, \mathcal{F}, \mathbb{P})$, which conditionally on $\mathcal{F}$, is a centred Gaussian martingale with conditional covariances

$$\widetilde{\mathbb{E}}(G_t G_t' | \mathcal{F}) = \boldsymbol{s}(\mathbf{j}, \mathbf{j}') \int_0^t \gamma_s^{q+q'} \mathrm{d}A_s, \quad \widetilde{\mathbb{E}}(H_t H_t' | \mathcal{F}) = \int_0^t \gamma_s^{q+q'} \overline{\alpha}_s \mathrm{d}A_s; \tag{C.16}$$
$$\widetilde{\mathbb{E}}(G_t H_t | \mathcal{F}) = \widetilde{\mathbb{E}}(G_t H_t' | \mathcal{F}) = \widetilde{\mathbb{E}}(G_t' H_t | \mathcal{F}) = \widetilde{\mathbb{E}}(G_t' H_t' | \mathcal{F}) = 0. \tag{C.17}$$

Since $\{u(\mathbf{j})_i^n\}_i$ are serially dependent, we will employ the "block splitting" technique that is often used in the literature (see, e.g., Jacod and Rosenbaum (2013) and Jacod et al. (2019)): we will divide the observations into "big blocks" of size pk_n separated by "small blocks" of size $K_{\mathbf{j},\mathbf{j}'} k_n$ where p will eventually grow to infinity and $K_{\mathbf{j},\mathbf{j}'}$ is a constant that depends on $\mathbf{j}, \mathbf{j}'$.

Now we consider small blocks of size $(2 + 2^{q-1})k_n$, and we need to introduce a sequence of notations associated with the block splitting techniques. By polarization, we will consider $\mathbf{j} = \mathbf{j}'$; moreover, $\boldsymbol{k}_n$ satisfies (14) thus is also fixed. We therefore write

θ_i^n instead of $\theta(\mathbf{j}, \boldsymbol{k}_n)_i^n$ in the sequel.

$$m(p,q) := p + 2 + 2^{q-1}, \quad J_n(p,t) := 1 + \left[\frac{n_t}{m(p,q)k_n}\right], \quad I_n(p,t) := q_n + J_n(p,t)m(p,q)k_n - 1;$$
$$\mathcal{H}_i^n := \mathcal{F}_i^n \otimes \mathcal{G}_{i-q_n-k_n+j_1}, \quad \mathcal{H}(p)_j^n := \mathcal{H}_{jm(p,q)k_n+q_n}^n, \quad \mathcal{H}'(p)_j^n := \mathcal{H}_{(jm(p,q)+p)k_n+q_n}^n;$$
$$\zeta(p)_i^n := \sum_{j=i}^{i+pk_n-1} \theta_j^n, \quad R(p)_t^n := \sum_{i=n_t-k_n-j_1+1}^{I_n(p,t)} \theta_i^n;$$
$$\eta(p)_j^n := \zeta(p)_{(j-1)m(p,q)k_n+q_n}^n, \quad \eta'(p)_j^n := \zeta(2+2^{q-1})_{((j-1)m(p,q)+p)k_n+q_n}^n;$$
$$\overline{\eta}(p)_j^n := \mathbb{E}\left(\eta(p)_j^n \,\middle|\mathcal{H}(p)_{j-1}^n\right), \quad \overline{\eta}'(p)_j^n := \mathbb{E}\left(\eta'(p)_j^n \,\middle|\mathcal{H}'(p)_{j-1}^n\right);$$
$$F(p)_t^n := \sum_{j=1}^{J_n(p,t)} \overline{\eta}(p)_j^n, \quad M(p)_t^n := \sum_{j=1}^{J_n(p,t)} (\eta(p)_j^n - \overline{\eta}(p)_j^n);$$
$$F'(p)_t^n := \sum_{j=1}^{J_n(p,t)} \overline{\eta}'(p)_j^n, \quad M'(p)_t^n := \sum_{j=1}^{J_n(p,t)} (\eta'(p)_j^n - \overline{\eta}'(p)_j^n).$$

Since $p \geq 2 + 2^{q-1}$, we conclude that $\eta(p)_j^n$ is $\mathcal{H}(p)_j^n$-measurable and $\eta'(p)_j^n$ is $\mathcal{H}'(p)_j^n$-measurable. Now it follows that

$$G_t^n = F(p)_t^n + F'(p)_t^n + M(p)_t^n + M'(p)_t^n - R(p)_t^n. \tag{C.18}$$

Lemma C.6. *For fixed $p \geq 2 + 2^{q-1}$, we have*

$$|\mathbb{E}\left(\zeta(p)_i^n \,|\mathcal{H}_i^n\right)| \leq K_p \delta_n{}^{\frac{1}{2}} k_n^{1-v}; \quad \left|\mathbb{E}\left((\zeta(p)_i^n)^4 \,|\mathcal{H}_i^n\right)\right| \leq K_p \delta_n^2 k_n^4.$$

Proof. By the independence of $\mathcal{G}, \mathcal{F}^{(0)}$, the boundedness of γ and Lemma B.1, we have for $j \geq i$ that

$$\left|\mathbb{E}\left(\theta_j^n \,|\mathcal{H}_i^n\right)\right| \leq K\sqrt{\delta_n}\left|\mathbb{E}\left(u(\mathbf{j})_j^n \,|\mathcal{G}_{i-q_n-k_n+j_1}\right)\right| \leq K\sqrt{\delta_n}(k_n + j - i)^{-v}.$$

Thus, we have $\left|\mathbb{E}\left(\zeta(p)_i^n \,|\mathcal{H}_i^n\right)\right| \leq K\sqrt{\delta_n}\sum_{j=i}^{i+pk_n-1}(k_n + j - i)^{-v} \leq K_p\sqrt{\delta_n}k_n^{1-v}$. The second estimate follows immediately from

$$\left|\mathbb{E}\left((\zeta(p)_i^n)^4 \,|\mathcal{H}_i^n\right)\right| \leq K \sum_{\ell_1 \leq \ell_2 \leq \ell_3 \leq \ell_4} \left|\mathbb{E}\left(\theta_{\ell_1}^n \theta_{\ell_2}^n \theta_{\ell_3}^n \theta_{\ell_4}^n \,|\mathcal{H}_i^n\right)\right| \leq K_p \delta_n^2 k_n^4.$$

This completes the proof. □

In the following lemma, we omit $\mathbf{j}$ and simply write $\boldsymbol{s}_\ell(k), \ell = 0, 1$, $\boldsymbol{s}_{k_n}(k)$ and $\boldsymbol{s}_{2,k_n}(k)$ instead of $\boldsymbol{s}_\ell(\mathbf{j}, \mathbf{j}; k), \ell = 0, 1$, $\boldsymbol{s}_{k_n}(\mathbf{j}, \mathbf{j}; k)$ and $\boldsymbol{s}_{2,k_n}(\mathbf{j}, \mathbf{j}; k)$.

Lemma C.7. *Let $v > 2$, for any $p \geq 2 + 2^{q-1}$, we have*

$$\left|\sum_{k=1}^{pk_n-1} \frac{k}{pk_n}\left(\boldsymbol{s}_{k_n}(k) + \boldsymbol{s}_{k_n}(-k)\right) - \frac{\mathfrak{C}_{\mathbf{j}}}{p}\right| \leq \frac{K_p}{k_n}, \tag{C.19}$$

where

$$\mathfrak{C}_{\mathbf{j}} := \sum_{(l,l'):j_l,j_{l'}\in\mathbf{j},l\neq l'} \boldsymbol{r}(\mathbf{j}_{-l})\boldsymbol{r}(\mathbf{j}_{-l'})\mathfrak{K}_{l,l'} - 2\sum_{l:j_l\in\mathbf{j}} \boldsymbol{r}(\mathbf{j}_{-l})\mathfrak{K}_l;$$

$$\mathfrak{K}_{l,l'} := \sum_{k=-\infty}^{\infty} \boldsymbol{r}(0,k) \times \begin{cases} 2^{l'-1}+1 & \textit{if } l=1, l'>1; \\ 2^{l-1}+1 & \textit{if } l'=1, l>1; \\ 2^{l\vee l'-1} - 2^{l\wedge l'-1} & \textit{if } l>1, l'>1, l\neq l'; \end{cases}$$

$$\mathfrak{K}_l := 2^{l-1}\sum_{k=-\infty}^{\infty} \boldsymbol{r}(\{0\}\oplus\mathbf{j}(+k)).$$

Proof. Since $v>2$, we have $\left|\sum_{k=1}^{pk_n-1}\frac{k}{pk_n}\left(\boldsymbol{s}_\ell(k)+\boldsymbol{s}_\ell(-k)\right)\right| \le K\left|\sum_{k=1}^{pk_n-1}\frac{k}{pk_n}\frac{1}{k^v}\right| \le \frac{K_p}{k_n}$ for $\ell=0,1$. Thus, it suffices to show

$$\left|\sum_{k=1}^{pk_n-1}\frac{k}{pk_n}\left(\boldsymbol{s}_{2,k_n}(k)+\boldsymbol{s}_{2,k_n}(-k)\right) - \frac{\mathfrak{C}_{\mathbf{j}}}{p}\right| \le \frac{K_p}{k_n}. \tag{C.20}$$

To see this, we will first show for $j_l, j_{l'}\in\mathbf{j}, l\neq l'$ (recall $\boldsymbol{r}_{k_n}$ defined in Lemma C.4)

$$\left|\sum_{k=1}^{pk_n-1}\frac{k}{pk_n}\left(\boldsymbol{r}_{k_n}(j_l,j_{l'}+k)+\boldsymbol{r}_{k_n}(j_l,j_{l'}-k)\right) - \frac{\mathfrak{K}_{l,l'}}{p}\right| \le \frac{K_p}{k_n}. \tag{C.21}$$

Let

$$k_n' := \begin{cases} (2^{l'-1}+1)k_n + j_l - j_{l'} & \text{if } l=1, l'>1; \\ (2^{l-1}+1)k_n + j_{l'} - j_l & \text{if } l'=1, l>1; \\ (2^{l\vee l'-1} - 2^{l\wedge l'-1})k_n + j_{l\wedge l'} - j_{l\vee l'} & \text{if } l>1, l'>1, l\neq l'. \end{cases}$$

Then (C.21) follows from

$$\sum_{k=1}^{pk_n-1}\left(\boldsymbol{r}_{k_n}(j_l,j_{l'}+k)+\boldsymbol{r}_{k_n}(j_l,j_{l'}-k)\right)k = \sum_{k=1-k_n'}^{pk_n-1-k_n'}\left(\boldsymbol{r}(0,k)+\boldsymbol{r}(0,k+2k_n')\right)(k+k_n'),$$

and the easy estimates that

$$\left|\sum_{k=1-k_n'}^{pk_n-1-k_n'}\boldsymbol{r}(0,k) - \sum_{k=-\infty}^{\infty}\boldsymbol{r}(0,k)\right| \le Kk_n^{1-v}, \quad \left|\sum_{k=1-k_n'}^{pk_n-1-k_n'}\boldsymbol{r}(0,k)k\right| \le K,$$

$$\left|\sum_{k=1-k_n'}^{pk_n-1-k_n'}\boldsymbol{r}(0,k+2k_n')(k+k_n')\right| \le Kk_n^{2-v}.$$

We can prove in a similar manner that for $j_l \in \mathbf{j}$,

$$\left|\sum_{k=1}^{pk_n-1} \frac{k}{pk_n}\left(\boldsymbol{r}_{k_n}(\{j_l+k\}\oplus\mathbf{j})+\boldsymbol{r}_{k_n}(\{j_l-k\}\oplus\mathbf{j})\right)-\frac{K_l}{p}\right| \le \frac{K_p}{k_n},$$
$$\left|\sum_{k=1}^{pk_n-1} \frac{k}{pk_n}\left(\boldsymbol{r}_{k_n}(\{j_l\}\oplus\mathbf{j}(+k))+\boldsymbol{r}_{k_n}(\{j_l\}\oplus\mathbf{j}(-k))\right)-\frac{K_l}{p}\right| \le \frac{K_p}{k_n}.$$

This finishes the proof of (C.20) and the proof is now complete.

□

Lemma C.8. *Let $v > 2$, for any $p \ge 2 + 2^{q-1}$, we have*

$$\left|\mathbb{E}\left((\zeta(p)_i^n)^2\,|\mathcal{H}_i^n\right) - pk_n\delta_n(\gamma_i^n)^{2q}\left(\boldsymbol{s}(\mathbf{j},\mathbf{j}) - \frac{\mathfrak{C}_\mathbf{j}}{p}\right)\right| \le K_pk_n\delta_n\left(k_n\delta_n^{\frac{1}{2}+\kappa} \vee k_n\delta_n^{\rho} \vee k_n^{-1}\right). \quad \text{(C.22)}$$

Proof. We have $(\zeta(p)_i^n)^2 = \sum_{j=i}^{\mu(p,0)_i^n}(\theta_j^n)^2 + 2\sum_{k=1}^{pk_n-1}\sum_{j=i}^{\mu(p,k)_i^n}\theta_j^n\theta_{j+k}^n$, where $\mu(p,k)_i^n := i-k+pk_n-1$. Thus, $\mathbb{E}\left((\zeta(p)_i^n)^2\,|\mathcal{H}_i^n\right) = \sum_{\ell=0}^{7}\mathfrak{E}(\ell)_{i,0}^{n,p} + 2\sum_{k=1}^{pk_n-1}\sum_{\ell=0}^{7}\mathfrak{E}(\ell)_{i,k}^{n,p}$, where for any nonnegative integer k

$$\begin{aligned}
\mathfrak{E}(0)_{i,k}^{n,p} &:= p\delta_nk_n(\gamma_i^n)^{2q}\boldsymbol{s}_{k_n}(k)\left(1-\frac{k}{pk_n}\right);\\
\mathfrak{E}(1)_{i,k}^{n,p} &:= pk_n\boldsymbol{s}_{k_n}(k)(\gamma_i^n)^{2q}\mathbb{E}\left(d_i^n\,|\mathcal{H}_i^n\right);\\
\mathfrak{E}(2)_{i,k}^{n,p} &:= -k\boldsymbol{s}_{k_n}(k)(\gamma_i^n)^{2q}\mathbb{E}\left(d_i^n\,|\mathcal{H}_i^n\right);\\
\mathfrak{E}(3)_{i,k}^{n,p} &:= \boldsymbol{s}_{k_n}(k)\sum\nolimits_{j=i}^{\mu(p,k)_i^n}\mathbb{E}\left((\gamma_j^n)^{2q}\alpha_j^n\delta(n,j+1)-(\gamma_i^n)^{2q}\alpha_i^n\delta(n,i+1)\,|\mathcal{H}_i^n\right);\\
\mathfrak{E}(4)_{i,k}^{n,p} &:= \boldsymbol{s}_{k_n}(k)\sum\nolimits_{j=i}^{\mu(p,k)_i^n}\mathbb{E}\left((\gamma_j^n)^q\left((\gamma_{j+k}^n)^q-(\gamma_j^n)^q\right)\alpha_j^n\delta(n,j+1)\,|\mathcal{H}_i^n\right);\\
\mathfrak{E}(5)_{i,k}^{n,p} &:= -\boldsymbol{s}_{k_n}(k)\sum\nolimits_{j=i}^{\mu(p,k)_i^n}\mathbb{E}\left((\gamma_j^n)^q(\gamma_{j+k}^n)^qd_j^n\,|\mathcal{H}_i^n\right);\\
\mathfrak{E}(6)_{i,k}^{n,p} &:= \sum\nolimits_{j=i}^{\mu(p,k)_i^n}\delta_n\left(\mathbb{E}\left(\theta_j^n\theta_{j+k}^n\,\middle|\mathcal{F}^{(0)}\right)-\boldsymbol{s}_{k_n}(k)\mathbb{E}\left((\gamma_j^n)^q(\gamma_{j+k}^n)^q\,|\mathcal{H}_i^n\right)\right);\\
\mathfrak{E}(7)_{i,k}^{n,p} &:= \sum\nolimits_{j=i}^{\mu(p,k)_i^n}\delta_n\left(\mathbb{E}\left(\theta_j^n\theta_{j+k}^n\,|\mathcal{H}_i^n\right)-\mathbb{E}\left(\theta_j^n\theta_{j+k}^n\,\middle|\mathcal{F}^{(0)}\right)\right).
\end{aligned}$$

First, we note by (A.2) that $\sum_{k=0}^{pk_n-1}\left|\mathfrak{E}(1)_{i,k}^{n,p}\right| \le K_p\delta_n^{\frac{3}{2}+\kappa}k_n$, and an application of Lemma C.7 yields a similar estimate $\sum_{k=0}^{pk_n-1}\left|\mathfrak{E}(2)_{i,k}^{n,p}\right| \le K_p\delta_n^{\frac{3}{2}+\kappa}k_n$. Next, we show

$$\sum_{k=0}^{pk_n-1}\left|\mathfrak{E}(3)_{i,k}^{n,p}\right| \le K_pk_n\delta_n(\delta_n^{\frac{1}{2}+\kappa} \vee k_n\delta_n^{\rho}). \quad \text{(C.23)}$$

Let $z(1)_{i,j}^n := \left((\gamma_j^n)^{2q}\alpha_j^n - (\gamma_i^n)^{2q}\alpha_i^n\right)\delta(n,j+1)$, $z(2)_{i,j}^n := (\gamma_i^n)^{2q}\alpha_i^n(\delta(n,j+1)-\delta(n,i+$

1)), then we have $\mathfrak{E}(3)^{n,p}_{i,k} = s_{k_n}(k)\sum_{j=i}^{\mu(p,k)^n_i}\mathbb{E}\left(z(1)^n_{i,j}+z(2)^n_{i,j}\,|\mathcal{H}^n_i\right)$. By first conditioning on $\mathcal{H}^n_i \vee \sigma(\delta(n,j+1))$, (A.2) and (A.2), we have $\left|\mathbb{E}\left(z(1)^n_{i,j}\,|\mathcal{H}^n_i\right)\right| \leq K(j-i)\delta_n^{1+\rho}$; similarly, we get $\left|\mathbb{E}\left((\alpha^n_i-\alpha^n_j)\delta(n,j+1)\,|\mathcal{H}^n_i\right)\right| \leq K(j-i)\delta_n^{1+\rho}$, together with the simple estimate (using again (A.2)) $\left|\mathbb{E}\left(\alpha^n_j\delta(n,j+1)-\alpha^n_i\delta(n,i+1)\,|\mathcal{H}^n_i\right)\right| \leq K\delta_n^{\frac{3}{2}+\kappa}$, we have

$$\left|\mathbb{E}\left(z(2)^n_{i,j}\,|\mathcal{H}^n_i\right)\right| \leq K\left((j-i)\delta_n^{1+\rho}\vee\delta_n^{3/2+\kappa}\right).$$

This proves (C.23).

Next, since for any $k>0$, $\delta(n,j+1)$ is independent of γ^n_{j+k} conditional on $\mathcal{H}^n_j$, we have by first conditioning on $\mathcal{H}^n_j$ and the estimate (A.1) that

$$\left|\mathbb{E}\left((\gamma^n_j)^q\left((\gamma^n_{j+k})^q-(\gamma^n_j)^q\right)\alpha^n_j\delta(n,j+1)\,|\mathcal{H}^n_i\right)\right| \leq K\delta_n^{1+\rho}k,$$

which implies $\sum_{k=0}^{pk_n-1}\left|\mathfrak{E}(4)^{n,p}_{i,k}\right| \leq K_pk_n^2\delta_n^{1+\rho}$. Similarly, we have $\sum_{k=0}^{pk_n-1}\left|\mathfrak{E}(5)^{n,p}_{i,k}\right| \leq K_pk_n^2\delta_n^{\frac{3}{2}+\kappa}$. Next, we can apply Lemma B.1 and Lemma C.5, which yields the following $\sum_{k=0}^{pk_n-1}\left(\left|\mathfrak{E}(6)^{n,p}_{i,k}\right|+\left|\mathfrak{E}(7)^{n,p}_{i,k}\right|\right) \leq K_p\delta_nk_n^{2-v}$. Now let $s(p)_{k_n} := \sum_{k=-(pk_n-1)}^{pk_n-1}s_{k_n}(k)$. We have

$$\begin{cases} |s(p)_{k_n}-s(\mathbf{j},\mathbf{j})| \leq K_pk_n^{1-v}, \\ \sum_{k=-(pk_n-1)}^{pk_n-1}\left(1-\frac{|k|}{pk_n}\right)s_{k_n}(k) = s(p)_{k_n}-\sum_{k=1}^{pk_n-1}\frac{k}{pk_n}\left(s_{k_n}(k)+s_{k_n}(-k)\right), \\ s_{2,k_n}(\mathbf{j},\mathbf{j};k) = s_{2,k_n}(\mathbf{j},\mathbf{j};-k). \end{cases}$$

The first estimate is obtained from Lemma C.7. Now we have

$$\left|\mathfrak{E}(0)^{n,p}_{i,0}+2\sum_{k=1}^{pk_n-1}\mathfrak{E}(0)^{n,p}_{i,k}-p\delta_nk_n(\gamma^n_i)^{2q}\left(s(\mathbf{j},\mathbf{j})-\frac{\mathfrak{C}_{\mathbf{j}}}{p}\right)\right| \leq K_p\delta_n.$$

This finishes the proof of the lemma. □

Lemma C.9. *If V is a càdlàg process, $p \geq 2+2^{q-1}$ and $\delta_n^\rho k_n \to 0$, we have for all $t>0$*

$$k_n\delta_n\sum_{j=1}^{J_n(p,t)}V_{(j-1)m(p,q)k_n+q_n} \xrightarrow{\mathbb{P}} \frac{\int_0^t V_s\mathrm{d}A_s}{m(p,q)}.$$

Proof. We only need to prove

$$k_n\delta_n\sum_{j=1}^{J_n(p,t)}V_{(j-1)m(p,q)k_n} \xrightarrow{\mathbb{P}} \frac{\int_0^t V_s\mathrm{d}A_s}{m(p,q)}, \tag{C.24}$$

since $k_n\delta_n \sum_{j=1}^{J_n(p,t)} \mathbb{E}\left(\left|V_{(j-1)m(p,q)k_n+q_n} - V_{(j-1)m(p,q)k_n}\right|\right) \leq K\sqrt{k_n\delta_n^\rho} \to 0$. Let $u_t^n := k_n\delta_n J_n(p,t)$. Then we have $k_n\delta_n \sum_{j=1}^{J_n(p,t)} V^n_{(j-1)m(p,q)k_n} = \int_0^{t+h_n} V_s \mathrm{d}u_s^n$ where $h_n := t^n_{J_n(p,t)m(p,q)k_n} - t^n_{n_t}$. It suffices to prove $h_n \xrightarrow{\mathbb{P}} 0$ since $u_t^n \xrightarrow{\mathbb{P}} A_t/m(p,q)$. Since $J_n(p,t)m(p,q)k_n - n_t \leq m(p,q)k_n$, we have for any $\epsilon > 0$

$$\limsup_n \mathbb{P}(|h_n| > \epsilon) \leq \limsup_n \mathbb{P}(A_{t+\epsilon} - A_t \leq m(p,q)k_n\delta_n) \to 0,$$

as $A_{t+\epsilon} > A_t, k_n\delta_n \to 0$. Now the proof of (C.24) is complete. □

Lemma C.10. *Let $\delta_n k_n^2 \to 0, v > 2, \delta_n k_n^{2v} \to \infty$ as $n \to \infty$, for all $p \geq 2 + 2^{q-1}$, we have*

$$\sum_{j=1}^{J_n(p,t)} \mathbb{E}\left((\eta(p)_j^n)^2 \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) \xrightarrow{\mathbb{P}} \frac{p}{m(p,q)} \left(\mathbf{s}(\mathbf{j},\mathbf{j}) - \frac{\mathfrak{C}_{\mathbf{j}}}{p}\right) \int_0^t \gamma_s^{2q} \mathrm{d}A_s; \tag{C.25}$$

$$F(p)_t^n \xrightarrow{\mathbb{P}} 0, \quad F'(p)_t^n \xrightarrow{\mathbb{P}} 0; \tag{C.26}$$

$$\mathbb{E}\left((M'(p)_t^n)^2\right) \leq \frac{K_t}{p}, \quad R(p)_t^n \xrightarrow{\mathbb{P}} 0. \tag{C.27}$$

Proof. (C.25) follows directly from Lemma C.8 and Lemma C.9. Since $J_n(p,t) \leq \frac{K_p t}{k_n\delta_n}$, we have by Lemma C.6 that $\mathbb{E}(|F(p)_t^n|) \leq \frac{K_p t}{\delta_n^{1/2} k_n^v} \to 0$; the same result applies to $F'(p)_t^n$. This proves (C.26). By the martingale property, we have

$$\mathbb{E}\left((M'(p)_t^n)^2\right) \leq \sum_{j=1}^{J_n(p,t)} \mathbb{E}\left(\left(\zeta(2+2^{q-1})^n_{((j-1)m(p,q)+p)k_n+q_n}\right)^2\right) \leq \frac{K_t}{p}.$$

The last inequality follows from Lemma C.8 and $J_n(p,t) \leq \frac{Kt}{p\delta_n k_n}$. Note that $I_n(p,t) - (n_t - k_n - j_1) \leq (p + 2(2^{q-1}+2))k_n$, therefore, we have

$$\mathbb{E}\left((R(p)_t^n)^2\right) \leq K_p k_n \sum_{i=n_t-k_n-j_1+1}^{n_t+(p+2(2^{q-1}+2))k_n} \mathbb{E}\left((\theta_i^n)^2\right) \leq K_p \delta_n k_n^2 \to 0.$$

This proves (C.27). □

Proposition C.1. *Let $v > 2, \delta_n k_n^3 \to 0$. For any fixed $p \geq 2 + 2^{q-1}$, the sequence of processes $M(p)^n$ converges $\mathcal{F}_\infty$-stably in law to the process $G(p)_t$ defined on an extension $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ of the original space, conditionally on $\mathcal{F}$, is centred Gaussian with (conditional) variance*

$$Z(p,\mathbf{j})_t := \frac{p}{m(p,q)} \left(\mathbf{s}(\mathbf{j},\mathbf{j}) - \frac{\mathfrak{C}_{\mathbf{j}}}{p}\right) \int_0^t \gamma_s^{2q} \mathrm{d}A_s.$$

Proof. Let $\widehat{\eta}(p)_j^n := \eta(p)_j^n - \overline{\eta}(p)_j^n$. Let $\Delta(V,p)_j^n := V^n_{jm(p,q)k_n+q_n} - V^n_{(j-1)m(p,q)k_n+q_n}$ for any process V. We also set $\mathcal{M} = \mathcal{M}_1 \cup W$, where W is the Brownian motion driving X and $\mathcal{M}_1$ denotes the class of all bounded $(\mathcal{F}_t)$-martingales orthogonal to W. By a standard stable convergence theorem for triangular arrays , see, e.g., Theorem IX 7.28 in Jacod and Shiryaev (2003), it suffices to prove the following three convergences:

$$\sum_{j=1}^{J_n(p,t)} \mathbb{E}\left((\widehat{\eta}(p)_j^n)^2 \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) \xrightarrow{\mathbb{P}} \frac{p}{m(p,q)}\left(s(\mathbf{j},\mathbf{j}) - \frac{\mathfrak{C}_{\mathbf{j}}}{p}\right)\int_0^t \gamma_s^{2q} \mathrm{d}A_s; \tag{C.28}$$

$$\sum_{j=1}^{J_n(p,t)} \mathbb{E}\left((\widehat{\eta}(p)_j^n)^4 \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) \xrightarrow{\mathbb{P}} 0; \tag{C.29}$$

$$\forall V \in \mathcal{M}, \quad \sum_{j=1}^{J_n(p,t)} \mathbb{E}\left(\widehat{\eta}(p)_j^n \Delta(V,p)_j^n \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) \xrightarrow{\mathbb{P}} 0. \tag{C.30}$$

(1) Note that $\mathbb{E}\left((\widehat{\eta}(p)_j^n)^2 \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) = \mathbb{E}\left((\eta(p)_j^n)^2 \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) - \mathbb{E}\left((\overline{\eta}(p)_j^n)^2 \,\middle|\, \mathcal{H}(p)_{j-1}^n\right)$, and from Lemma C.6, we have $(\overline{\eta}(p)_j^n)^2 \le K_p \delta_n k_n^{2-2v}$. Since $J_n(p,t) \le \frac{K_p t}{\delta_n k_n}$, we conclude that

$$\sum_{j=1}^{J_n(p,t)} \mathbb{E}\left((\overline{\eta}(p)_j^n)^2 \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) \le K_p k_n^{1-2v} \to 0.$$

Now (C.28) follows from the first part of Lemma C.10.

(2) By Lemma C.6, we have

$$\sum_{j=1}^{J_n(p,t)} \mathbb{E}\left((\eta(p)_j^n)^4 \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) \le K_p \delta_n k_n^3, \qquad \sum_{j=1}^{J_n(p,t)} \mathbb{E}\left((\overline{\eta}(p)_j^n)^4 \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) \le K_p \delta_n k_n^{3-4v}.$$

Now (C.29) is proved.

(3) It suffices to show

$$\sum_{j=1}^{J_n(p,t)} \mathbb{E}\left(\eta(p)_j^n \Delta(V,p)_j^n \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) \xrightarrow{\mathbb{P}} 0, \tag{C.31}$$

since $\mathbb{E}\left(\Delta(V,p)_j^n \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) = 0$ for any $V \in \mathcal{M}$. Consider any i in the range $[(j-1)m(p,q)k_n + q_n, jm(p,q) - 2k_n - 1]$. We have $\mathbb{E}\left(\theta_i^n \Delta(V,p)_j^n \,\middle|\, \mathcal{H}(p)_{j-1}^n\right) = \mathfrak{X}(1)_{i,j}^n \mathfrak{X}(2)_{i,j}^n$, where

$$\mathfrak{X}(1)_{i,j}^n := \mathbb{E}\left(\sqrt{\delta_n}(\gamma_i^n)^q \Delta(V,p)_j^n \,\middle|\, \mathcal{F}_{(j-1)m(p,q)k_n+q_n}^n\right), \; \mathfrak{X}(2)_{i,j}^n := \mathbb{E}\left(u(\mathbf{j})_i^n \,\middle|\, \mathcal{G}_{((j-1)m(p,q)-1)k_n+j_1}\right).$$

By Lemma B.1, we have $\sum_i \left|\mathfrak{X}(2)_{i,j}^n\right| \le K_p k_n^{1-v}$; now we apply the boundedness of γ

and get $\left|\mathfrak{X}(1)^n_{i,j}\right| \leq K\sqrt{\delta_n \mathbb{E}\left((\Delta(V,p)^n_j)^2 \left|\mathcal{F}^n_{(j-1)m(p,q)k_n+q_n}\right.\right)}$. Thus, we have

$$\left|\mathbb{E}\left(\eta(p)^n_j \Delta(V,p)^n_j \left|\mathcal{H}(p)^n_{j-1}\right.\right)\right| \leq K_p k_n^{1-v}\sqrt{\delta_n \mathbb{E}\left((\Delta(V,p)^n_j)^2 \left|\mathcal{F}^n_{(j-1)m(p,q)k_n+q_n}\right.\right)},$$

and an application of the Cauchy-Schwarz inequality and the martingale property of V yield

$$\begin{aligned}\mathbb{E}\left(\left(\sum_{j=1}^{J_n(p,t)} \mathbb{E}\left(\eta(p)^n_j \Delta(V,p)^n_j \left|\mathcal{H}(p)^n_{j-1}\right.\right)\right)^2\right) &\leq K_{p,t} k_n^{1-2v} \mathbb{E}\left(\sum_{j=1}^{J_n(p,t)} (\Delta(V,p)^n_j)^2\right) \\ &\leq K_{p,t} k_n^{1-2v} \mathbb{E}\left((V_{t^n_{J_n(p,t)m(p,q)k_n+q_n}} - V_0)^2\right).\end{aligned}$$

If $V \in \mathcal{M}_1$, $\mathbb{E}\left((V_{t^n_{J_n(p,t)m(p,q)k_n+q_n}} - V_0)^2\right)$ is further bounded by $\mathbb{E}\left((V_\infty - V_0)^2\right) < K$. This proves (C.31) with $V \in \mathcal{M}_1$. When $V = W$, $t^n_{J_n(p,t)m(p,q)k_n+q_n} \leq t+1$ for n large enough on the set Ω^n_t (recall (A.3)). Thus, (C.31) is proved with $V = W$ on the set Ω^n_t. Since $\mathbb{P}(\Omega^n_t) \to 1$, the proof is complete for $V = W$.

□

Theorem C.2. *$(M(p)^n_t, H^n_t)$ convergences $\mathcal{F}_\infty$-stably in law to $(G(p)_t, H_t)$ that is defined on an extension $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ of $(\Omega, \mathcal{F}, \mathbb{P})$, which conditionally on $\mathcal{F}$, is a centred Gaussian martingale with conditional covariances*

$$\widetilde{\mathbb{E}}(G(p)_t G(p)_t|\mathcal{F}) = Z(p,\mathbf{j})_t, \quad \widetilde{\mathbb{E}}(H_t H_t'|\mathcal{F}) = \int_0^t \gamma_s^{q+q'} \overline{\alpha}_s \mathrm{d}A_s, \quad \widetilde{\mathbb{E}}(G(p)_t H_t|\mathcal{F}) = 0.$$

Proof. Lemma C.8 yields an estimate that $\mathbb{E}\left((\widehat{\eta}(p)^n_j)^2\right) \leq K_p \delta_n k_n$. Now we have $\mathbb{E}\left(\sum_{j=1}^{J_n(p,t)} \mathbb{E}\left(\left(\widehat{\eta}(p)^n_j\right)^2 \mathbf{1}_{\{\widehat{\eta}(p)^n_j > \epsilon\}} \left|\mathcal{H}(p)^n_{j-1}\right.\right)\right) \to 0$ by Lebesgue's dominated convergence theorem and the fact that $\delta_n k_n \to 0$. This in turn leads to the following convergence for any $\epsilon > 0$

$$\sum_{j=1}^{J_n(p,t)} \mathbb{E}\left(\left(\widehat{\eta}(p)^n_j\right)^2 \mathbf{1}_{\{\widehat{\eta}(p)^n_j > \epsilon\}} \left|\mathcal{H}(p)^n_{j-1}\right.\right) \xrightarrow{\mathbb{P}} 0. \tag{C.32}$$

On the other hand, we have $e^{\mathrm{i}uM(p)^n_t} = g(u,p)^n_t \mathfrak{M}(u,p)^n_t$, $e^{\mathrm{i}uG(p)_t} = g(u,p)_t \mathfrak{M}(u,p)_t$ where $g(u,p)^n_t$ and $g(u,p)_t$ are predictable with finite variation, and $\mathfrak{M}(u,p)^n_t$ and $\mathfrak{M}(u,p)_t$ are martingales (see, e.g., Theorem II.2.47 in Jacod and Shiryaev (2003)). According to the proof of Theorem VIII.2.4 in Jacod and Shiryaev (2003) (see also the proof of Theorem A.4 in Jacod et al. (2017)), (C.28) and (C.32) imply $g(u,p)^n_t \xrightarrow{\mathbb{P}}$

$g(u,p)_t$. Now the joint convergence follows from Proposition C.1 and Theorem A.4 of Jacod et al. (2017). □

Proof of Theorem C.1. By polarization, it suffices to consider $\mathbf{j} = \mathbf{j}'$. The process $V(p)^n := G^n_t - M(p)^n$ satisfies $\lim_{p\to\infty} \limsup_{n\to\infty} \mathbb{P}(|V(p)^n_t| > \epsilon) = 0$ for all $\epsilon > 0, t > 0$. This follows from (C.18) and Lemma C.10. On the other hand, $Z(p,\mathbf{j})_t(\omega) \le K$ and $Z(p,\mathbf{j})_t(\omega) \to Z(\mathbf{j},\mathbf{j})_t(\omega)$ for all $t > 0$ and ω, we thus have $G(p)_t \xrightarrow{\mathbb{P}} G_t$. Now Theorem C.1 follows from Theorem C.2. □

The next lemma will be used to prove the consistency of the proposed estimators for the asymptotic variances and covariances. The first part of the lemma (C.33) is similar to but more general than Lemma B.3.

Lemma C.11. *Let $\mathbf{j}_\ell \in \mathfrak{J}$, $\mathbf{j}_\ell = (j_{\ell,1},\dots,j_{\ell,q_\ell})$, $q_\ell = |\mathbf{j}_\ell|, \ell = 1,2,\dots,d$. $\{w^n_\ell\}_{\ell=1}^d$ is a sequence of integers satisfying $w_{n,1} = 0$, $w^n_{\ell+1} - w^n_\ell \ge 2^{q_{\ell+1}-1}k_n + j_{\ell,1} + 2k_n$ for $\ell \ge 1$. Let $\overline{w}^n_d := w^n_d \vee k_n \vee \bar{j}$, where $\bar{j} := \max\{j_{\ell,p} : 1 \le \ell \le d, 1 \le p \le q_\ell\}$. Assume $\delta_n^\rho \overline{w}^n_d \to 0$ and $v > 1$. Let*

$$\mathfrak{U}^n_{d,t} := \sum_{i=2^{q_1-1}k_n}^{n_t-(j_{d,1}+w^n_d\vee k_n)} \prod_{\ell=1}^d \Delta_{\mathbf{j}_\ell}(Y)^n_{i+w^n_\ell}; \quad \mathfrak{U}'^n_{d,t} := \sum_{i=2^{q_1-1}k_n}^{n_t-(j_{d,1}+w^n_d\vee k_n)} (\gamma^n_i)^{\bar{q}} \prod_{\ell=1}^d \Delta_{\mathbf{j}_\ell}(\chi)^n_{i+w^n_\ell};$$
$$\mathfrak{U}''^n_{d,t} := \sum_{i=2^{q_1-1}k_n}^{n_t-(j_{d,1}+w^n_d\vee k_n)} (\gamma^n_i)^{\bar{q}} \left(\prod_{\ell=1}^d \Delta_{\mathbf{j}_\ell}(\chi)^n_{i+w^n_\ell} - \prod_{\ell=1}^d \boldsymbol{r}(\mathbf{j}_\ell;k_n) \right).$$

Then we have

$$\mathbb{E}\left(\left| \mathfrak{U}^n_{d,t} - \mathfrak{U}'^n_{d,t} \right| \mathbf{1}_{\{\Omega^n_t\}} \right) \le K_r (\overline{w}^n_d)^{1/r} \delta_n^{\rho/r-1}, \tag{C.33}$$
$$\mathbb{E}\left(\left(\mathfrak{U}''^n_{d,t} \right)^2 \mathbf{1}_{\{\Omega^n_t\}} \right) \le K \left(\overline{w}^n_d \delta_n^{-1} + (\delta_n k_n^v)^{-2} \right). \tag{C.34}$$

Proof. Let $s_0 = 0, s_\ell := s_{\ell-1} + q_\ell$, $\ell \ge 1$. Let $\{\iota_l\}_{l=1}^{\bar{q}}$ be an enumeration of $\{j_{\ell,p} : 1 \le \ell \le d, 1 \le p \le q_\ell\}$ such that $\iota_l = j_{\ell,l-s_{\ell-1}}$ if $s_{\ell-1} < l \le s_\ell$. That is, for each $1 \le l \le \bar{q}$, there is a unique pair $(\ell(l), p(l))$ such that $\iota_l = j_{\ell(l),p(l)}$.

Let $\zeta^n_{i,m,m'} := X^n_{i+m} - X^n_{i+m'} + (\gamma^n_{i+m} - \gamma^n_i)\chi_{i+m} - (\gamma^n_{i+m'} - \gamma^n_i)\chi_{i+m'}$, $\zeta'^n_{i,m,m'} := \gamma^n_i(\chi_{i+m} - \chi_{i+m'})$. For any integer $p \ge 1$, we let $k_{p,n} = -k_n$ if $p = 1$ and $k_{p,n} = 2^{p-1}k_n$ if $p > 1$. Now let $m_l := w^n_{\ell(l)} + j_{\ell(l),p(l)}$; $m'_l := w^n_{\ell(l)} + j_{\ell(l),p(l)} - k_{p(l),n}$. Using the notations, we obtain (recall (B.3) for $\mathcal{Q}_{\overline{q}}$)

$$\begin{aligned}&\prod_{\ell=1}^{d}\Delta_{\mathbf{j}_\ell}(Y)^n_{i+w^n_\ell}=\prod_{l=1}^{\overline{q}}(\zeta^n_{i,m_l,m'_l}+\zeta'^n_{i,m_l,m'_l}),\quad (\gamma^n_i)^{\overline{q}}\prod_{\ell=1}^{d}\Delta_{\mathbf{j}_\ell}(\chi)^n_{i+w^n_\ell}=\prod_{l=1}^{\overline{q}}\zeta'^n_{i,m_l,m'_l};\\&\prod_{\ell=1}^{d}\Delta_{\mathbf{j}_\ell}(Y)^n_{i+w^n_\ell}-(\gamma^n_i)^{\overline{q}}\prod_{\ell=1}^{d}\Delta_{\mathbf{j}_\ell}(\chi)^n_{i+w^n_\ell}=\sum_{Q\in\mathcal{Q}_{\overline{q}}}\prod_{l\in Q}\zeta'^n_{i,m_l,m'_l}\prod_{l\in Q^c}\zeta^n_{i,m_l,m'_l}.\end{aligned}\tag{C.35}$$

Apply (A.1) for X and γ, and the fact that χ has bounded moments, we get for any $k\geq 2$ that

$$\mathbb{E}\left(|\zeta^n_{i,m_l,m'_l}|^k\right)\leq K\delta^\rho_n\overline{w}^n_d;\quad \mathbb{E}\left(|\zeta'^n_{i,m,m'}|^k\right)\leq K.$$

For a fixed $Q\in\mathcal{Q}_{\overline{q}}$, let $\mu=|Q^c|$ whence $\mu\geq 1$. For $r\geq 2$, apply the Hölder's inequality with exponents $(\underbrace{r\mu,\ldots,r\mu}_{\mu},\frac{r}{r-1})$, we get

$$\mathbb{E}\left(\left|\prod_{l\in Q}\zeta'^n_{i,m_l,m'_l}\prod_{l\in Q^c}\zeta^n_{i,m_l,m'_l}\right|\right)\leq\prod_{l\in Q^c}\left(\mathbb{E}\left(\left|\zeta^n_{i,m_l,m'_l}\right|^{r\mu}\right)\right)^{\frac{1}{r\mu}}\left(\mathbb{E}\left(\left|\prod_{l\in Q}\zeta'^n_{i,m_l,m'_l}\right|^{\frac{r}{r-1}}\right)\right)^{\frac{r-1}{r}}\leq K(\delta^\rho_n\overline{w}^n_d)^{1/r}.\tag{C.36}$$

For $1<r\leq 2$, we note (C.36) still holds. Now let's consider $Q^c=\{l^*\}$. Let (ℓ^*,p^*) be the associated pair such that $\iota_{l^*}=j_{\ell^*,p^*}$. Let $\mathcal{H}^n_{i,l^*}:=\mathcal{F}^n_{i+m'_{l^*}}\otimes\mathcal{G}$ if $p^*=1$, and $\mathcal{H}^n_{i,l^*}:=\mathcal{F}^n_{i+w_{n,\ell^*}}\otimes\mathcal{G}$ if $p^*>1$. Thus, we have $\left|\mathbb{E}\left(\zeta^n_{i,m_{l^*},m'_{l^*}}\middle|\mathcal{H}^n_{i,l^*}\right)\right|\leq K\delta^\rho_n\overline{w}^n_d(1+|\chi_{i+m_{l^*}}|+|\chi_{i+m'_{l^*}}|)$, which yields $\mathbb{E}\left(\left|\mathbb{E}\left(\zeta^n_{i,m_{l^*},m'_{l^*}}\prod_{l\neq l^*}\zeta'^n_{i,m_l,m'_l}\middle|\mathcal{H}^n_{i,l^*}\right)\right|\right)\leq K\delta^\rho_n\overline{w}^n_d$ since

$$\mathbb{E}\left(\left|\mathbb{E}\left(\zeta^n_{i,m_{l^*},m'_{l^*}}\prod_{l\neq l^*}\zeta'^n_{i,m_l,m'_l}\,|\mathcal{H}^n_{i,l^*}\right)\right|\right)\leq K\overline{w}^n_d\delta^\rho_n\mathbb{E}\left((1+|\chi_{i+m_{l^*}}|+|\chi_{i+m'_{l^*}}|)\prod_{l\neq l^*}\left|\chi_{i+m_l}-\chi_{i+m'_l}\right|\right).$$

On the other hand, we have by Hölder's inequality (since $r>1$) that

$$\mathbb{E}\left((\zeta^n_{i,m_{l^*},m'_{l^*}}\prod_{l\neq l^*}\zeta'^n_{i,m_l,m'_l})^2\right)\leq K(\delta^\rho_n(\overline{w}^n_d))^{1/r}.$$

Also note that $\zeta^n_{i,m_{l^*},m'_{l^*}}\prod_{l\neq l^*}\zeta'^n_{i,m_l,m'_l}$ is measurable with respect to $\mathcal{F}^n_{i+w_{n,\ell^*}+j_{\ell^*,1}-k_{1,n}}\otimes\mathcal{G}$, we thus have by Lemma A.6 in Jacod et al. (2017) that

$$\begin{aligned}\mathbb{E}\left(\left|\sum_{i=2^{q_1-1}k_n}^{n_t-(j_{d,1}+\overline{w}^n_d)k_n}\zeta^n_{i,m_{l^*},m'_{l^*}}\prod_{l\neq l^*}\zeta'^n_{i,m_l,m'_l}\right|\mathbf{1}_{\{\Omega^n_i\}}\right)&\leq K_r\left(\overline{w}^n_d\delta^{\rho-1}_n+(\overline{w}^n_d)^{(r+1)/2r}\delta^{(\rho-r)/2r}_n\right)\\&\leq K_r(\overline{w}^n_d)^{1/r}\delta^{\rho/r-1}_n.\end{aligned}$$

The last inequality is due to $\delta_n^\rho \overline{w}_d^n \to 0$. This proves (C.33).

Now we prove (C.34). Let $\varpi_i^n := \prod_{\ell=1}^d \Delta_{\mathbf{j}_\ell}(\chi)_{i+w_\ell^n}^n - \prod_{\ell=1}^d \boldsymbol{r}(\mathbf{j}_\ell; k_n)$ (When the index set is empty, we let the product be 1, e.g., for $\ell = 1$, $\prod_{\ell'=1}^{\ell-1} \Delta_{\mathbf{j}_{\ell'}}(\chi)_{i+w_{n,\ell'}}^n = 1$). Then $\varpi_i^n = \sum_{\ell=1}^d \varpi_{i,\ell}^n$, where $\varpi_{i,\ell}^n := u(\mathbf{j}_\ell)_{i+w_\ell^n}^n \prod_{\ell'=1}^{\ell-1} \Delta_{\mathbf{j}_{\ell'}}(\chi)_{i+w_{n,\ell'}}^n \prod_{\ell''=\ell+1}^d \boldsymbol{r}(\mathbf{j}_{\ell''}; k_n)$. By Lemma B.1, we have $\left|\mathbb{E}\left(\varpi_{i,\ell}^n\right)\right| \le K k_n^{-v}$. Next, using again Lemma B.1, we have for any $1 \le \ell, \ell' \le d$ that

$$\left|\mathbb{E}\left(\varpi_{i,\ell}^n \varpi_{i+l,\ell'}^n\right)\right| \le K k_n^{-2v} + K((l - h_n) \vee 1)^{-v}, \tag{C.37}$$

where $h_n := \overline{w}_d^n + \bar{j} + (2^{q_1 - 1} + 1) k_n$. Since γ is bounded and $v > 1$, we have (C.34) by (C.37). □

Lemma C.12. *Assume all conditions of Theorem 4.3 hold. We have*

$$\frac{\widehat{\sigma}_1(\mathbf{j}, \mathbf{j}')_t^n}{n_t} \xrightarrow{\mathbb{P}} \frac{\boldsymbol{s}(\mathbf{j}, \mathbf{j}')}{A_t} \int_0^t \gamma_s^{q''} \mathrm{d}A_s. \tag{C.38}$$

Proof. For $\ell = 0, 1, 2$, let $S_\ell(\mathbf{j}, \mathbf{j}'; i_n)_1 := s_\ell(\mathbf{j}, \mathbf{j}'; 0) + \sum_{k=1}^{i_n} (s_\ell(\mathbf{j}, \mathbf{j}'; k) + s_\ell(\mathbf{j}', \mathbf{j}; k))$, and $S_\ell(\mathbf{j}, \mathbf{j}'; i_n)_2 := \sum_{k > i_n} (s_\ell(\mathbf{j}, \mathbf{j}'; k) + s_\ell(\mathbf{j}', \mathbf{j}; k))$. Let $\boldsymbol{s}_\ell(\mathbf{j}, \mathbf{j}') := S_\ell(\mathbf{j}, \mathbf{j}'; i_n)_1 + S_\ell(\mathbf{j}, \mathbf{j}'; i_n)_2$. We first prove

$$\begin{aligned}&\delta_n \left(U(7, 0; \mathbf{j}, \mathbf{j}')_t^n + \sum_{k=1}^{i_n} \left(U(7, k; \mathbf{j}, \mathbf{j}')_t^n + U(7, k; \mathbf{j}', \mathbf{j})_t^n \right) + (2 i_n + 1) U(4, \mathbf{j}, \mathbf{j}')_t^n \right) \\ &\xrightarrow{\mathbb{P}} \boldsymbol{s}_0(\mathbf{j}, \mathbf{j}') \int_0^t \gamma_s^{q''} \mathrm{d}A_s.\end{aligned} \tag{C.39}$$

Since $v > 1$, γ is bounded, we have $S_\ell(\mathbf{j}, \mathbf{j}'; i_n)_2 \int_0^t \gamma_s^{q''} \mathrm{d}A_s \le K i_n^{1-v} \to 0$. It is therefore sufficient to replace $\boldsymbol{s}_0(\mathbf{j}, \mathbf{j}') \int_0^t \gamma_s^{q''} \mathrm{d}A_s$ by $S_\ell(\mathbf{j}, \mathbf{j}'; i_n)_1 \int_0^t \gamma_s^{q''} \mathrm{d}A_s$ on the RHS of (C.39). Using the decomposition (D.1) (for $\mathbf{j} \oplus \mathbf{j}'(+k)$), for $k \le i_n$, we have $\mathbb{E}(\delta_n (G_t^n)^2) \le K \delta_n (k_n \vee i_n)$; Lemma B.3 gives $\mathbb{E}\left(\left|\sqrt{\delta_n} \mathfrak{R}(\mathbf{j} \oplus \mathbf{j}'(+k), 2)\right|\right) \le K_r (\delta_n^\rho (k_n \vee i_n))^{1/r}$, since $\rho > 1/2$, we can find some $r > 1$ such that $(\delta_n^\rho (k_n \vee i_n))^{1/r} \le \sqrt{\delta_n (k_n \vee i_n)}$ whence $\mathbb{E}\left(\left|\sqrt{\delta_n} \mathfrak{R}(\mathbf{j} \oplus \mathbf{j}'(+k), 2)\right|\right) \le K \sqrt{\delta_n (k_n \vee i_n)}$; we also have $\mathbb{E}\left(\left|\sqrt{\delta_n} \mathfrak{R}(\mathbf{j} \oplus \mathbf{j}'(+k), 3)\right|\right) \le K k_n^{-v}$ by Lemma B.1; since $i_n^2 \delta_n \to 0$, we have by Lemma A.2 and Lemma A.7 in Jacod et al. (2017) that $(|H_t^n| + |\mathfrak{R}(\mathbf{j} \oplus \mathbf{j}'(+k), 1)|) \sqrt{\delta_n} i_n \xrightarrow{\mathbb{P}} 0$ since $(k_n \vee i_n) \delta_n^{\rho - 1/2} \to 0$. Therefore, we have

$$\sum_{k=1}^{i_n} \left(\delta_n U(7, k; \mathbf{j}, \mathbf{j}')_t^n - \boldsymbol{r}(\mathbf{j} \oplus \mathbf{j}'(+k)) \int_0^t \gamma_s^{q''} \mathrm{d}A_s \right) \xrightarrow{\mathbb{P}} 0. \tag{C.40}$$

Let $U'(4;\mathbf{j},\mathbf{j}')^n_t := -\sum_{i=2^{q-1}k_n}^{n_t-w(4)_n} (\gamma^n_i)^{q''} \Delta_{\mathbf{j}}(\chi)^n_i \Delta_{\mathbf{j}'}(\chi)^n_{i+w(4)^n_2}$, $C_4(\mathbf{j},\mathbf{j}')^n_t := \delta_n U(4,\mathbf{j},\mathbf{j}')^n_t + \boldsymbol{r}(\mathbf{j})\boldsymbol{r}(\mathbf{j}') \int_0^t \gamma_s^{q''} \mathrm{d}A_s$. $C_4(\mathbf{j},\mathbf{j}')^n_t$ can be decomposed into

$$C_4(\mathbf{j},\mathbf{j}')^n_t = \sum_{\ell=1}^{5} \mathfrak{D}_4(\ell)^n_t, \tag{C.41}$$

where

$$\begin{aligned}
\mathfrak{D}_4(1)^n_t &:= \delta_n (U(4,\mathbf{j},\mathbf{j}')^n_t - U'(4,\mathbf{j},\mathbf{j}')^n_t); \\
\mathfrak{D}_4(2)^n_t &:= (\boldsymbol{r}(\mathbf{j})\boldsymbol{r}(\mathbf{j}') - \boldsymbol{r}(\mathbf{j};k_n)\boldsymbol{r}(\mathbf{j}';k_n)) \int_0^t \gamma_s^{q''} \mathrm{d}A_s; \\
\mathfrak{D}_4(3)^n_t &:= \boldsymbol{r}(\mathbf{j};k_n)\boldsymbol{r}(\mathbf{j}';k_n) \left(\int_0^t \gamma_s^{q''} \mathrm{d}A_s - \sum_{i=2^{q-1}k_n}^{n_t-w(4)_n} (\gamma^n_i)^{q''} \alpha^n_i \delta(n,i+1) \right); \\
\mathfrak{D}_4(4)^n_t &:= \boldsymbol{r}(\mathbf{j};k_n)\boldsymbol{r}(\mathbf{j}';k_n) \sum_{i=2^{q-1}k_n}^{n_t-w(4)_n} (\gamma^n_i)^{q''} (\alpha^n_i \delta(n,i+1) - \delta_n); \\
\mathfrak{D}_4(5)^n_t &:= \delta_n \sum_{i=2^{q-1}k_n}^{n_t-w(4)_n} (\gamma^n_i)^{q''} \left(\Delta_{\mathbf{j}}(\chi)^n_i \Delta_{\mathbf{j}'}(\chi)^n_{i+w(4)^n_2} - \boldsymbol{r}(\mathbf{j};k_n)\boldsymbol{r}(\mathbf{j}';k_n) \right).
\end{aligned}$$

Now we'll prove

$$i_n \left(\delta_n U(4,\mathbf{j},\mathbf{j}')^n_t + \boldsymbol{r}(\mathbf{j})\boldsymbol{r}(\mathbf{j}') \int_0^t \gamma_s^{q''} \mathrm{d}A_s \right) \xrightarrow{\mathbb{P}} 0 \tag{C.42}$$

by almost repeating the analysis to obtain (C.40): $\mathbb{E}\left(|\mathfrak{D}_4(1)^n_t| \mathbf{1}_{\{\Omega^n_t\}} \right) \leq K_r (\delta_n^{\rho} k_n)^{1/r}$ by Lemma C.11; since χ has bounded moments of all orders and γ is bounded, Lemma B.2 leads to $|\mathfrak{D}_4(2)^n_t| \leq K k_n^{-v}$; next, Lemma A.2 and Lemma A.7 in Jacod et al. (2017) imply $\sqrt{\delta_n}(\mathfrak{D}_4(3)^n_t + \mathfrak{D}_4(4)^n_t) i_n \xrightarrow{\mathbb{P}} 0$; a direct application of the second part of Lemma C.11 gives $\mathbb{E}\left((\mathfrak{D}_4(5)^n_t)^2 \right) \leq K(\delta_n k_n + \delta_n k_n^{-2v})$. This proves (C.42) and together with (C.40), we have (C.39).

We can prove in a similar manner

$$\delta_n \left(U(5,0;\mathbf{j},\mathbf{j}')^n_t + \sum_{k=1}^{i_n} (U(5,k;\mathbf{j},\mathbf{j}')^n_t + U(5,k;\mathbf{j}',\mathbf{j})^n_t) \right) \xrightarrow{\mathbb{P}} \boldsymbol{s}_1(\mathbf{j},\mathbf{j}') \int_0^t \gamma_s^{q''} \mathrm{d}A_s. \tag{C.43}$$

$$\delta_n \left(U(6,0;\mathbf{j},\mathbf{j}')^n_t + \sum_{k=1}^{i_n} (U(6,k;\mathbf{j},\mathbf{j}')^n_t + U(6,k;\mathbf{j}',\mathbf{j})^n_t) \right) \xrightarrow{\mathbb{P}} \boldsymbol{s}_2(\mathbf{j},\mathbf{j}') \int_0^t \gamma_s^{q''} \mathrm{d}A_s. \tag{C.44}$$

(C.38) follows from (4), (C.39), (C.43) and (C.44). □

Lemma C.13. *Assume all conditions of Theorem 4.3 hold. We have*

$$\frac{1}{n_t}\widehat{\sigma}_2(\mathbf{j},\mathbf{j}')^n_t \xrightarrow{\mathbb{P}} \frac{\boldsymbol{r}(\mathbf{j})\boldsymbol{r}(\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q''}\overline{\alpha}_s \mathrm{d}A_s; \tag{C.45}$$

$$\frac{1}{n_t}\widehat{\sigma}_3(\mathbf{j},\mathbf{j}')^n_t \xrightarrow{\mathbb{P}} \frac{\mathbf{R}(\mathbf{j})\mathbf{R}(\mathbf{j}')}{A_t}\int_0^t \overline{\alpha}_s \mathrm{d}A_s - \frac{\mathbf{R}(\mathbf{j})\boldsymbol{r}(\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q'}\overline{\alpha}_s \mathrm{d}A_s - \frac{\boldsymbol{r}(\mathbf{j})\mathbf{R}(\mathbf{j}')}{A_t}\int_0^t \gamma_s^{q}\overline{\alpha}_s \mathrm{d}A_s. \tag{C.46}$$

Proof. We prove (C.45), and

$$\frac{U(\mathbf{j})^n_t}{n_t} \xrightarrow{\mathbb{P}} \frac{\boldsymbol{r}(\mathbf{j})}{A_t}\int_0^t \gamma_s^{q}\overline{\alpha}_s \mathrm{d}A_s \tag{C.47}$$

can be proved analogously. In view of (3.17) in Jacod et al. (2017), (C.46) follows immediately from (C.47).

We first introduce the following notations:

$$B^n_t := \delta_n U(3,\mathbf{j},\mathbf{j}')^n_t - \sum_{i=q_n}^{n_t - w(3)_n} \delta_n \boldsymbol{r}(\mathbf{j};k_n)\boldsymbol{r}(\mathbf{j}';k_n)\overline{\alpha}^n_i(\gamma^n_i)^{q''}; \quad \widetilde{\mathcal{H}}^n_i := \mathcal{F}^n_i \otimes \mathcal{G}_{i+2+2k_n};$$

$$\mathfrak{B}(1)^n_i := \widehat{\delta}^n_i\left(\Delta_{\mathbf{j}}(Y)^n_{i+w(3)^n_2}\Delta_{\mathbf{j}'}(Y)^n_{i+w(3)^n_3} - (\gamma^n_i)^{q''}\Delta_{\mathbf{j}}(\chi)^n_{i+w(3)^n_2}\Delta_{\mathbf{j}'}(\chi)^n_{i+w(3)^n_3}\right);$$

$$\mathfrak{B}(2)^n_i := \widehat{\delta}^n_i(\gamma^n_i)^{q''}\left(\Delta_{\mathbf{j}}(\chi)^n_{i+w(3)^n_2}\Delta_{\mathbf{j}'}(\chi)^n_{i+w(3)^n_3} - \boldsymbol{r}(\mathbf{j};k_n)\boldsymbol{r}(\mathbf{j}';k_n)\right);$$

$$\mathfrak{B}(3)^n_i := \boldsymbol{r}(\mathbf{j};k_n)\boldsymbol{r}(\mathbf{j}';k_n)(\gamma^n_i)^{q''}(\widehat{\delta}^n_i - \overline{\alpha}^n_i).$$

Then we have an easy estimate by the independence of $\mathcal{F}^{(0)}$ and $\mathcal{G}$ and Lemma B.1 that $\mathbb{E}\left(\left|\mathbb{E}\left(\mathfrak{B}(2)^n_i\middle|\widetilde{\mathcal{H}}^n_i\right)\right|\right) \leq Kk_n^{-v}$. Moreover, we have $\mathbb{E}\left((\mathfrak{B}(2)^n_i)^2\right) \leq K$ since $\mathbb{E}\left((\widehat{\delta}^n_i)^2\,\middle|\mathcal{F}^n_i\right) \leq K$ (see the proof of Lemma A.10 in Jacod et al. (2017)). Since $\mathfrak{B}(2)^n_i$ is $\widetilde{\mathcal{H}}^n_{i+q_n+q'_n+3k_n+j_1}$-measurable. Now Lemma A.6 in Jacod et al. (2017) yields $\mathbb{E}\left(\left|\sum_{i=0}^{n_t-w(3)_n}\mathfrak{B}(2)^n_i\right|\mathbf{1}_{\{\Omega^n_t\}}\right) \leq Kk_n^{1/2}\delta_n^{-1/2}$. Using the decomposition (C.35), apply Hölder's inequality, we obtain

$$\mathbb{E}\left(\left(\Delta_{\mathbf{j}}(Y)^n_{i+w(3)^n_2}\Delta_{\mathbf{j}'}(Y)^n_{i+w(3)^n_3} - (\gamma^n_i)^{q''}\Delta_{\mathbf{j}}(\chi)^n_{i+w(3)^n_2}\Delta_{\mathbf{j}'}(\chi)^n_{i+w(3)^n_3}\right)^2\right) \leq K_r(\delta_n^{\rho}k_n)^{1/r},$$

Apply Cauchy-Schwarz inequality, we have $\mathbb{E}\left(\left|\sum_{i=q_n}^{n_t-w(3)_n}\mathfrak{B}(1)^n_i\right|\mathbf{1}_{\{\Omega^n_t\}}\right) \leq K_{r,q}k_n^{\frac{1}{2r}}\delta_n^{\frac{\rho}{2r}-1}$. Next, we have by Lemma A.10 in Jacod et al. (2017) that $\mathbb{E}\left(\left|\sum_{i=q_n}^{n_t-w(3)_n}\mathfrak{B}(3)^n_i\right|\mathbf{1}_{\{\Omega^n_t\}}\right) \leq \frac{K}{\delta_n k_n}$. Since $|B^n_t| \leq \delta_n\left|\sum_{\ell=1}^{3}\sum_{i=q_n}^{n_t-w(3)_n}\mathfrak{B}(\ell)^n_i\right|$ and $\mathbb{P}(\Omega^n_t) \to 1$, we have $B^n_t \xrightarrow{\mathbb{P}} 0$. Now the proof of (C.45) is complete. □

D Proof of the Main Theorems

Let

$$Z(\mathbf{j})_t^n =: G_t^n - \boldsymbol{r}(\mathbf{j};k_n)H_t^n + \sum_{\ell=1}^{3} \mathfrak{R}(\mathbf{j},\ell)_t^n, \tag{D.1}$$

where

$$\mathfrak{R}(\mathbf{j},1)_t^n := -\frac{\boldsymbol{r}(\mathbf{j};\boldsymbol{k}_n)}{\sqrt{\delta_n}} \left(\int_0^t \gamma_s^q \mathrm{d}A_s - \sum_{i=q_n}^{n_t-k_n-j_1} (\gamma_i^n)^q \alpha_i^n \delta(n,i+1) \right);$$
$$\mathfrak{R}(\mathbf{j},2)_t^n := \sqrt{\delta_n} \left(\mathrm{ReMeDI}(Y;\mathbf{j},\boldsymbol{k}_n)_t^n - \mathrm{ReMeDI}'(\chi;\mathbf{j},\boldsymbol{k}_n)_t^n \right);$$
$$\mathfrak{R}(\mathbf{j},3)_t^n := \frac{\boldsymbol{r}(\mathbf{j};\boldsymbol{k}_n) - \boldsymbol{r}(\mathbf{j})}{\sqrt{\delta_n}} \int_0^t \gamma_s^q \mathrm{d}A_s.$$

Proof of Theorem 4.1. It suffices to show $\sqrt{\delta_n} Z(\mathbf{j})_t^n \xrightarrow{\mathbb{P}} 0$ in view of (4). Since γ is bounded, a direct application of Lemma B.1 yields $\mathbb{E}\left(\delta_n (G_t^n)^2\right) \le K\delta_n \bar{k}_n \to 0$; Lemma A.2 and Lemma A.7 of Jacod et al. (2017) imply $\sqrt{\delta_n}\left(|\boldsymbol{r}(\mathbf{j};k_n)H_t^n| + |\mathfrak{R}(\mathbf{j},1)_t^n|\right) \xrightarrow{\mathbb{P}} 0$; Lemma B.3 yields $\sqrt{\delta_n}\mathbb{E}\left(|\mathfrak{R}(\mathbf{j},2)_t^n|\,\mathbf{1}_{\{\Omega_t^n\}}\right) \le K(\delta_n^\rho \bar{k}_n)^{1/r} \to 0$ whence $\sqrt{\delta_n}\mathfrak{R}(\mathbf{j},2)_t^n \xrightarrow{\mathbb{P}} 0$ since $\mathbb{P}(\Omega_t^n) \to 1$; Lemma B.2 gives $\sqrt{\delta_n}\,|\mathfrak{R}(\mathbf{j},3)_t^n| \to 0$. This completes the proof. □

Proof of Theorem 4.2. Now we have the following:

$$\mathfrak{R}(\mathbf{j},1)_t^n \xrightarrow{\mathbb{P}} 0; \quad \mathbb{E}\left(|\mathfrak{R}(\mathbf{j},2)_t^n|\,\Big|\mathbf{1}_{\{\Omega_t^n\}}\right) \le K k_n^{1/r} \delta_n^{\rho/r-1/2}; \quad |\mathfrak{R}(\mathbf{j},3)_t^n| \le K k_n^{-v} \delta_n^{-1/2};$$

which are immediate in view of Lemma A.7 in Jacod et al. (2017), Lemma B.3 and Lemma B.2, respectively. Thus, we have $\mathfrak{R}(\mathbf{j},2)_t^n \xrightarrow{\mathbb{P}} 0$ for r close to 1 and $\mathbb{P}(\Omega_t^n) \to 1$, and $\mathfrak{R}(\mathbf{j},3)_t^n \xrightarrow{\mathbb{P}} 0$ in view of (14). Now the first part of Theorem 4.2 is a simple consequence of Theorem C.1 and part (b) follows directly the proof of Theorem 3.4 in Jacod et al. (2017). □

Proof of Theorem 4.3. The convergence is an immediate result of Theorem 4.2, Lemma C.12 and Lemma C.13. □

Supplementary Materials for "A ReMeDI for Microstructure Noise"

Additional Materials (Not for Publication)

Z. Merrick Li[§] Oliver Linton[¶]

September 8, 2020

E Additional Simulation Studies

Recall the settings of the simulation studies

$$\mathrm{d}X_t = \kappa_1(\mu_1 - X_t)\mathrm{d}t + \sigma_t \mathrm{d}W_{1,t} + \xi_{1,t}\mathrm{d}N_t; \quad \mathrm{d}\sigma_t^2 = \kappa_2(\mu_2 - \sigma_t^2)\mathrm{d}t + \eta\sigma_t \mathrm{d}W_{2,t} + \xi_{2,t}\mathrm{d}N_t;$$
$$\mathbf{Corr}(W_1, W_2) = v; \quad \xi_{1,t} \sim \mathcal{N}\left(0, \mu_2/10\right); N_t \sim \mathrm{Poi}(\lambda); \xi_{2,t} \sim \mathrm{Exp}(\delta),$$

where we set

$$\kappa_1 = 0.5;\ \mu_1 = 3.6;\ \kappa_2 = 5/252;\ \mu_2 = 0.04/252;\ \eta = 0.05/252;\ v = -0.5;\ \lambda = 1;\ \delta = \eta.$$

For the noise process:

$$\chi_{i+1} = \rho\chi_i + e_i, \quad e_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, 1-\rho^2\right), \quad |\rho| < 1,$$
$$\gamma_t = C_\gamma \gamma_t', \quad \mathrm{d}\gamma_t' = -\rho_\gamma(\gamma_t' - \mu_t)\mathrm{d}t + \sigma_\gamma \mathrm{d}W_t.$$

We set $\rho = 0.7$, $\rho_\gamma = 10$, $\mu_t = 1 + 0.1\cos(2\pi t)$, $\sigma_\gamma = 0.1$, $C_\gamma = 5 \times 10^{-4}$. For the observation times: $\{t_i^n\}_i$ follow an inhomogeneous Poisson process with rate $n\alpha_t$ where $\alpha_t = (1 + \cos(2\pi t))/2$.

[§]Faculty of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom. Email: z.merrick.li@gmail.com

[¶]Faculty of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom. Email: obl20@cam.ac.uk.

E.1 The feasible CLT

We examine the feasible CLT presented in (21) of Li and Linton (2020) for $\ell = 0, 1, 2$. The QQ-plots in Figure E.1 clearly support the limit distribution.

E.2 Discreteness of prices

This section studies the "discreteness" of asset prices. With the minimal ticks, the observed price becomes $[Y_i^n]$ instead of Y_i^n, where $[Y_i^n]$ is the rounded price. Consequently, the "noise", the deviation of the observed price from the efficient price, has an additional component due to the discreteness:

$$\varepsilon_i^n = [Y_i^n] - X_i^n = \underbrace{[Y_i^n] - Y_i^n}_{\text{discreteness error}} + \underbrace{Y_i^n - X_i^n}_{\text{microstructure noise}} . \tag{E.1}$$

In our simulation design, we round the observed real price to 1 cent,

$$[Y_i^n] = \log\left([100 \exp\left(Y_i^n\right)]/100\right). \tag{E.2}$$

In Figure E.2, we plot the mean estimates (the blue solid line) of the first 20 autocorrelations of the rounded noise based on the observed prices $[Y_i^n]$ in the left panel. Compared to the true parameters (red stars), we see that the ReMeDI estimator retains its accuracy in the presence of additional error due to price discreteness. In the right panel of Figure E.2, we estimate the autocorrelations functions using the sample autocorrelation function of noise $\varepsilon_i^n = [Y_i^n] - X_i^n$ *directly*. As expected, the estimates are more accurate than the ReMeDI estimates. But they are not feasible in practice as noise is latent.

E.3 Robustness to the choices of tuning parameters

Both ReMeDI and LA are nonparametric estimators and the tuning parameter k_n plays an important role in the real performance. To study the robustness to the choices of k_n, we select a wide range of tuning parameters for both estimators and plot the mean-squared error (MSE) in Figure E.3 for each k_n. We observe that the LA estimators are quite sensitive to the value of k_n, and the MSE increases as k_n increases. The ReMeDI estimators, however, remains small MSE across the different k_ns. Thus, we conclude that the ReMeDI approach is very robust to the choice of tuning parameters compared to the LA method.

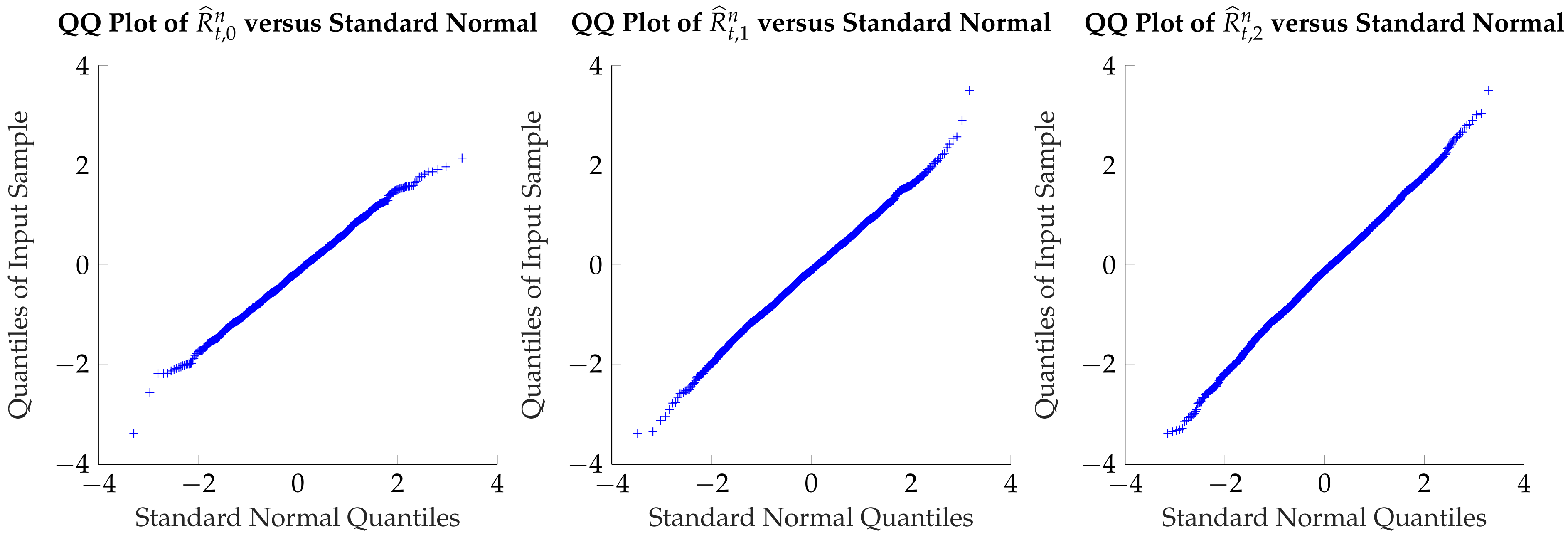

Figure E.1: Standard normal QQ-plots of the ReMeDI estimators $\widehat{R}^n_{t,\ell}$ defined in (19) for $\ell = 0, 1, 2$. The turning parameters are $k_n = 15, i_n = 5, n = 46,800, \phi_n = k_n^{3/5}/n$. The number of simulation is 1,000.

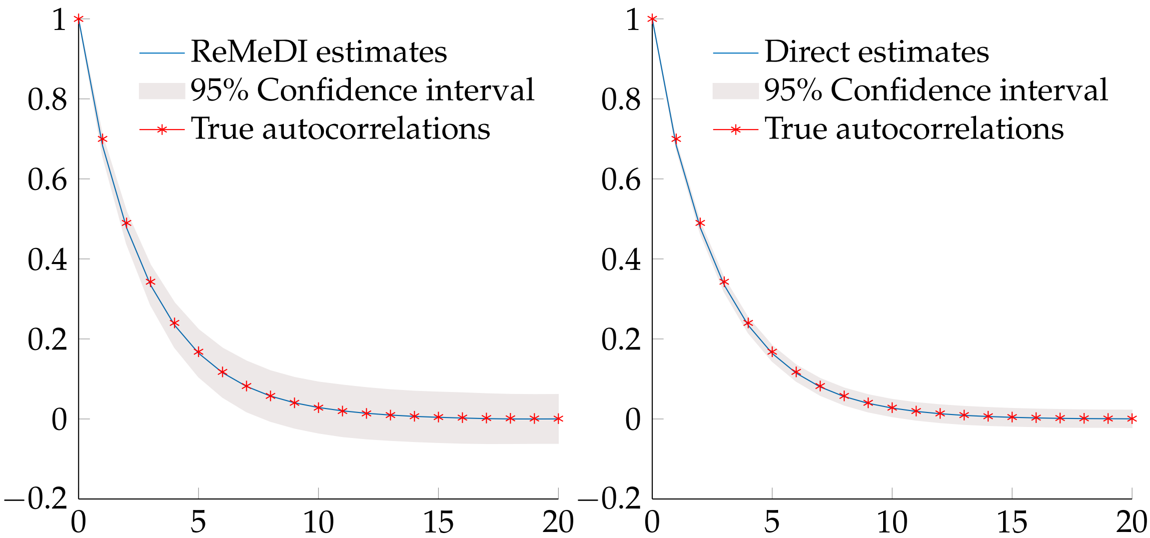

Figure E.2: Estimation of the autocorrelation function of noise using rounded price. The blue solid line in the left panel is the mean estimates of 1,000 simulations by the ReMeDI estimators based on rounded observed price. The blue solid line in the right panel is the mean estimates of 1,000 simulations by the sample autocovariance function based on the noise directly. The shaded areas are the simulated 95% confidence intervals. The red stars are the true parameters. We set $n = 23,400$. The tuning parameter is $k_n = 10$.

E.4 Simulation in the Jacod et al. (2017) setting

This section replicates the simulation studies of Jacod et al. (2017), in which the stationary part of the noise χ follows a moving average process with strong and persistent autocorrelation patterns.[1] It is also interesting to examine the performance of the ReMeDI estimators in this setting. Moreover, we perturb some key parameters to assess the robustness of the two approaches.

Figure E.4 reports the estimates of autocovariances of noise in exactly the same setting as in Jacod et al. (2017). It seems that the LA approach outperforms the ReMeDI estimators. However, Figure E.5 shows that the specific choice of σ (the constant volatility of the efficient price) and the tuning parameter of LA leads to the outperformance. In the top panel, we see that the advantage of LA over ReMeDI disappears when the tuning parameter varies slightly, and LA performs less well when the tuning parameter becomes larger. The bottom panel reveals that performance of LA largely depends on the volatility; a small perturbation of the volatility leads to large biases and variances. But the ReMeDI approach is relatively more robust.

[1] Jacod et al. (2017) claim the simulation setting is motivated by their empirical findings; however, as we have demonstrated via our simulation studies that the extremely strong autocorrelation patterns found are largely due to the finite sample bias, see another study by Li et al. (2020).

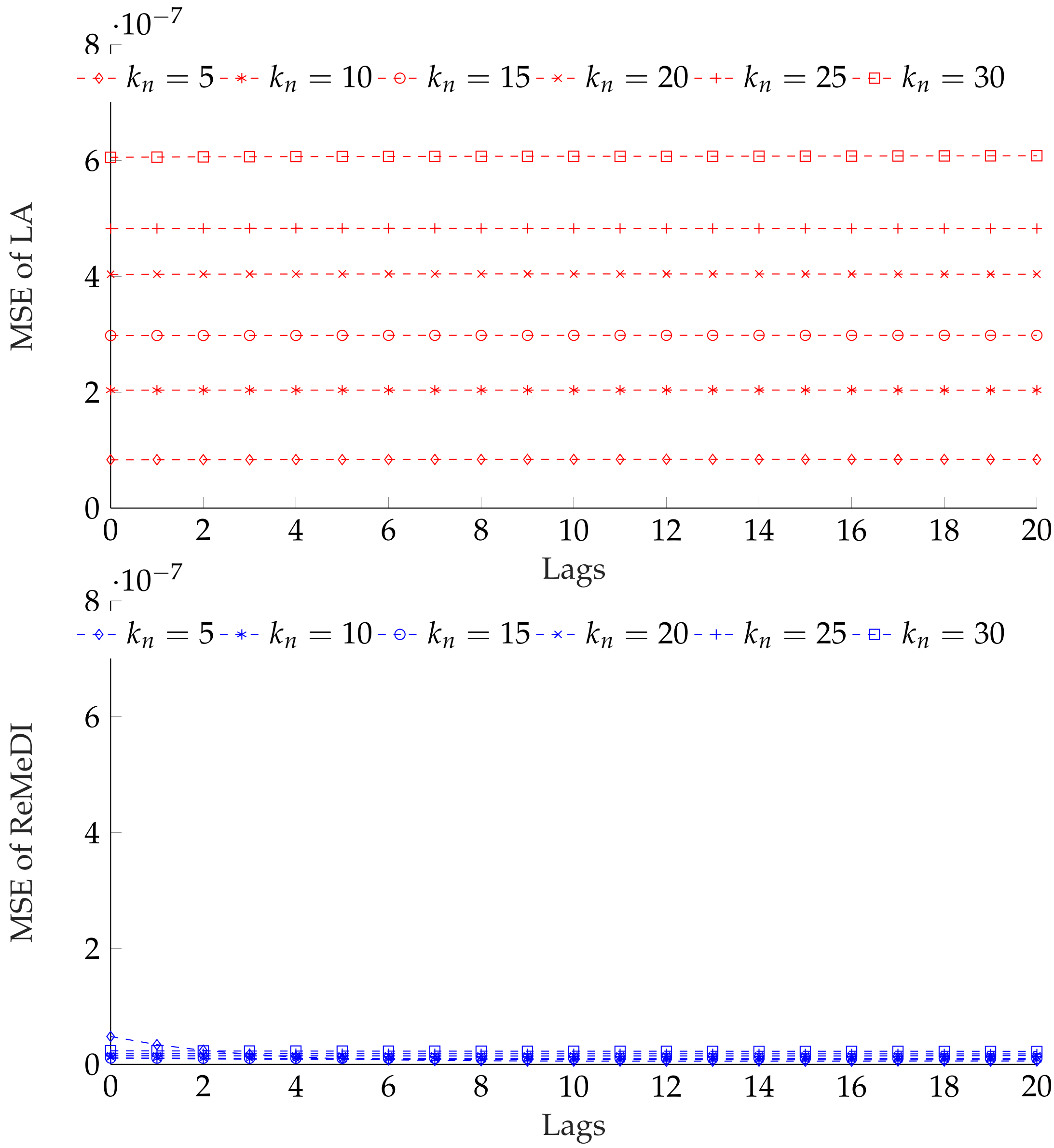

Figure E.3: Mean squared error (MSE) of ReMeDI and LA estimators of the first 20 autocovariances of noise with different choices of the tuning parameter. We set $\gamma_t' \equiv 1 \, \forall t$ so that the noise process is stationary. We consider $k_n = 5, 10, 15, 20, 25, 30$. We set $n = 23,400$.

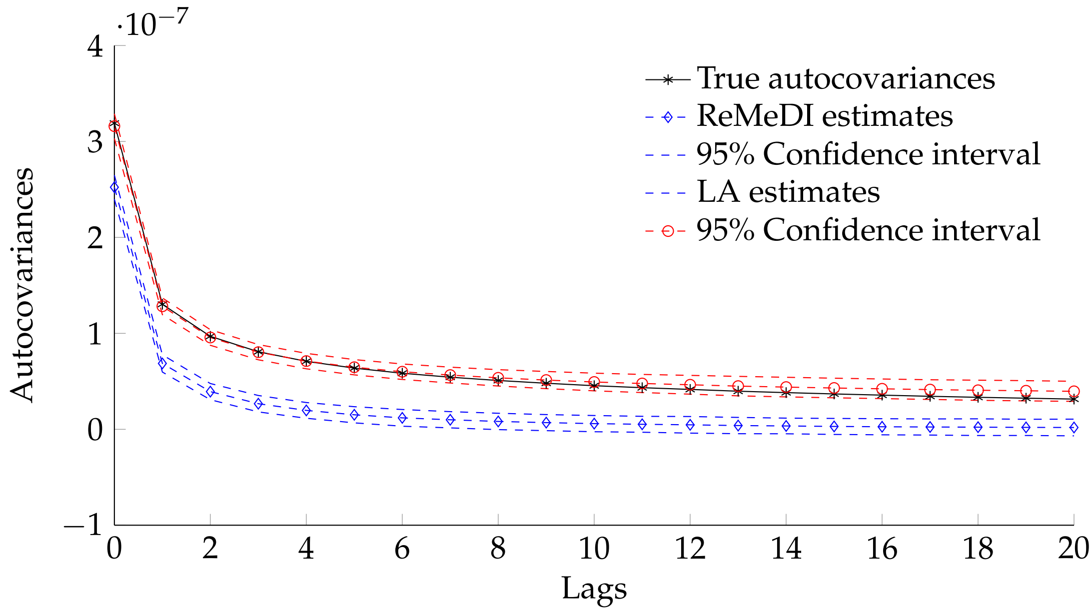

Figure E.4: Estimating autocovariances of noise using the simulation setting in Jacod et al. (2017). The black stars are the mean estimates of the sample autocovariances of noise based on 1,000 simulations. The blue-dotted-diamonds are the mean ReMeDI estimates, and the dotted blue lines are the 95% confidence intervals. The red-dotted-circles are the mean LA estimates and the dotted red lines are the 95% confidence intervals. The tuning parameter for the ReMeDI estimators is fixed at $k_n = 10$.

F On the Choice of the Tuning Parameter k_n

Like many nonparametric estimators, the implementation of the ReMeDI estimators needs to select some tuning parameter, which controls the lengths of the disjoint intervals here. We propose a simple rule to choose the tuning parameter. We construct a simple statistic as the benchmark. Its probability limit is a function of several true moments of the noise. We then estimate those moments by ReMeDI for a given tuning parameter and compare with the benchmark statistics. This gives an "error function" for a selected tuning parameter, and we select the one with a small error. The selection procedure is able to adjust the choices of the tuning parameters to reflect the degree of serial dependence: it selects a smaller (or larger) tuning parameter when the serial dependence in the microstructure noise is weaker (or stronger).

F.1 Methodology 1

We assume $\boldsymbol{k_n}$ satisfies (14) in Li and Linton (2020). The methodology introduced here can also be applied to other choices of $\boldsymbol{k_n}$, e.g., $\boldsymbol{k_n}$ specified in (12).

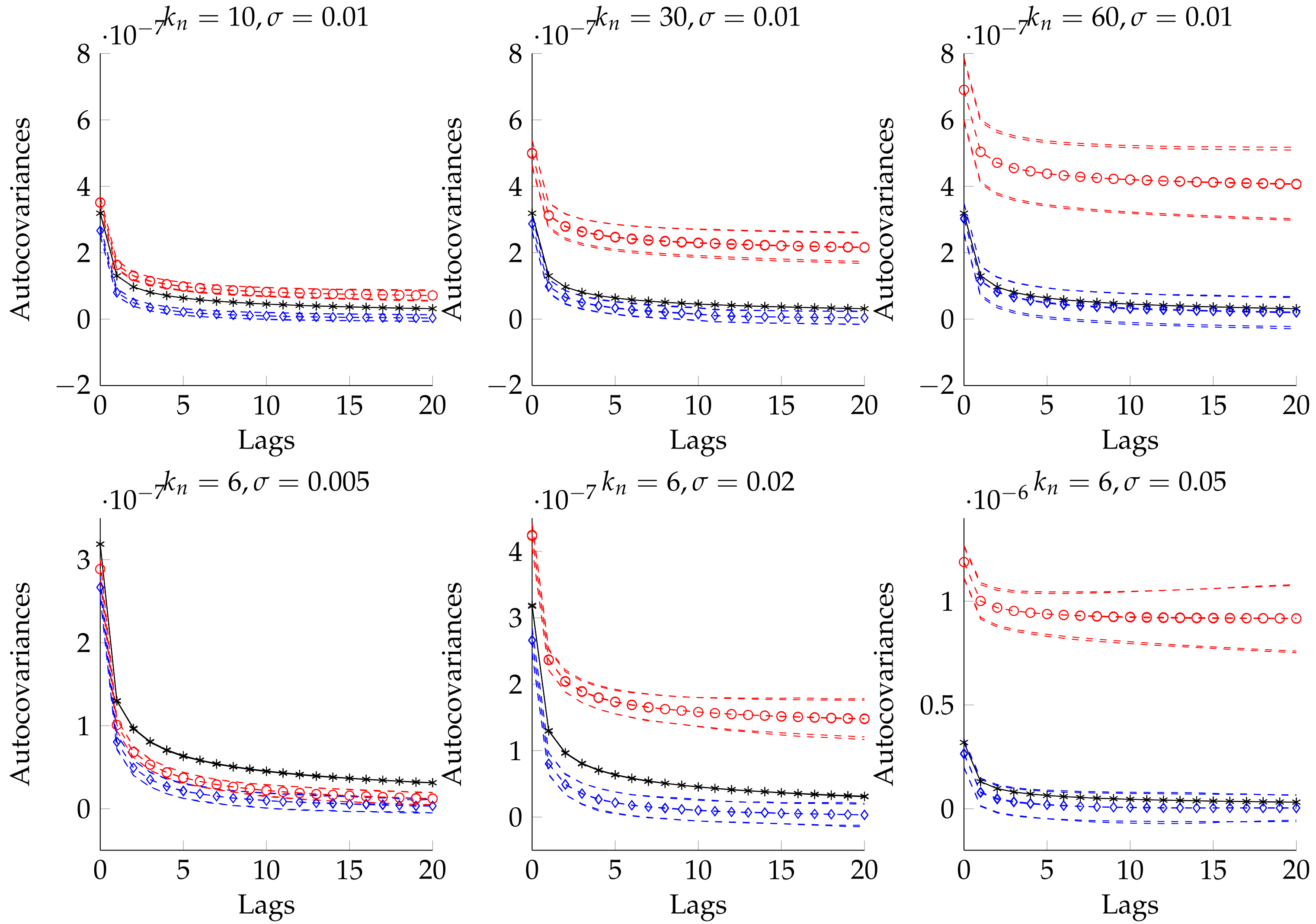

Figure E.5: Estimating autocovariances of noise using the simulation setting in Jacod et al. (2017). The black stars are the mean estimates of the sample autocovariances of noise based on 1,000 simulations. The blue-dotted-diamonds are the mean ReMeDI estimates, and the dotted blue lines are the 95% confidence intervals. The red-dotted-circles are the mean LA estimates and the dotted red lines are the 95% confidence intervals. In the top panel, the constant volatility is fixed at $\sigma = 0.01$; the tuning parameter for both ReMeDI and LA are 10, 30, and 60; respectively. In the bottom panel, the tuning parameter is fixed at $k_n = 6$ while the volatility varies from 0.005, to 0.02 and 0.05.

Let $\boldsymbol{k}_1 := (-1, 2)$ and $\mathbf{j}_0 := (0, 0)$, we obtain the following estimates

$$\frac{1}{n_t}\text{ReMeDI}(Y; \mathbf{j}_0, \boldsymbol{k}_1)_t^n \xrightarrow{\mathbb{P}} \frac{\int_0^t \gamma_s^2 \mathrm{d}s}{A_t}\left(r_0 - r_1 - r_2 + r_3\right). \tag{F.1}$$

$\text{ReMeDI}(Y; \mathbf{j}_0, \boldsymbol{k}_1)_t^n / n_t$ provides a precise proxy of the probability limit for at least three reasons. First, the statistics are free of tuning parameters; second, the "non-overlapping intervals" trick effectively removes any bias due to the efficient price; third, the non-overlapping intervals appear in $\text{ReMeDI}(Y; \mathbf{j}_0, \boldsymbol{k}_1)_t^n / n_t$ are very short, thus the variation induced by the efficient price is largely reduced. Therefore, we consider $\text{ReMeDI}(Y; \mathbf{j}_0, \boldsymbol{k}_1)_t^n / n_t$ as a "good" yet simple proxy that compresses the information of the first 4 autocovariances of the microstructure noise.

We rewrite $\widehat{R}_{t,\ell}^n$ (defined in (19) of Li and Linton (2020)) as $\widehat{R}_{t,\ell}^{n,k_n}$ to highlight the dependence of $\widehat{R}_{t,\ell}^n$ on k_n. Given arbitrary k_n, we obtain $\widehat{R}_{t,\ell}^{n,k_n}, \ell = 0, 1, 2, 3$, as the estimates of the first 4 autocovariances of the noise. Now we define a random variable as a measurement of the squared estimation errors:

$$\text{SqErr}(k_n)_t^n := \left(\widehat{R}_{t,0}^{n,k_n} - \widehat{R}_{t,1}^{n,k_n} - \widehat{R}_{t,2}^{n,k_n} + \widehat{R}_{t,3}^{n,k_n} - \frac{1}{n_t}\text{ReMeDI}(Y; \mathbf{j}_0, \boldsymbol{k}_1)_t^n\right)^2. \tag{F.2}$$

Intuitively, $\text{SqErr}(k_n)_t^n$ measures (the square of) the bias of estimating $\int_0^t \gamma_s^2 \mathrm{d}s (r_0 - r_1 - r_2 + r_3) / A_t$ for a given k_n. It is decreasing in k_n if noise is autocorrelated since larger k_n helps to reduce the bias. Figure F.1 plots the mean estimates of $\text{SqErr}(k_n)_t^n$ against different k_n using simulated data. Indeed, we observe fast decreasing curves. Therefore, the plots suggest that larger k_n is always preferred to minimize the bias. However, larger k_n increases the size of the intervals, inducing more variation caused by the efficient price.

Now the selection rule is clear: select the "minimal" k_n at which the error function $\text{SqErr}(k_n)_t^n$ is perceived as "close" to zero. Consider an example, in which the noise follows an AR(1) process with the AR(1) coefficient $\rho = 0.7$. The plot in the top panel of Figure F.1 suggests that proper choices of k_n would be $k_n = 8, 9$, or 10. When the serial dependence becomes weaker, a smaller k_n would be sufficient to reduce the bias. The selection rule will take the degree of the dependence into account and acts accordingly. We observe in the bottom panel of Figure F.1 that when $\rho = 0.4$, the selection rule will pick $k_n = 3, 4$, or 5.

Algorithm 1 Choose k_n

$k_n^{\max} \leftarrow$ the maximum choice of k_n
$\delta_{\text{tol}} \leftarrow$ error tolerance
$d_{\text{lsw}} \leftarrow$ size of local searching window
$\mathbf{Err} \leftarrow$ error vector
$[l, u] \leftarrow$ simple search range

```
%Compute the error function (F.2) for each k in range
```

for $k = 1$ to $k_n^{\max} + 1 + d_{\text{lsw}}$ **do**

  $\widehat{R}_{t,0}^{n,k}, \widehat{R}_{t,1}^{n,k}, \widehat{R}_{t,2}^{n,k}, \widehat{R}_{t,3}^{n,k} \leftarrow$ ReMeDI with k
  $\mathbf{Err}(k) \leftarrow \text{SqErr}(k)_t^n$

$\text{Err}^{\max} \leftarrow \max(\mathbf{Err}(1 : \lfloor k_n^{\max}/2 \rfloor))$

```
%Identify the ''flat'' part of the error function with small values
```

for $\tilde{k} = 1$ to $k_n^{\max} + 1$ **do**

  $\text{LocalErr}^{\min} \leftarrow \min(\mathbf{Err}(\tilde{k} : \tilde{k} + d_{\text{lsw}}))$
  $\text{LocalErr}^{\max} \leftarrow \max(\mathbf{Err}(\tilde{k} : \tilde{k} + d_{\text{lsw}}))$

  if $\max\{\text{LocalErr}^{\min}/\text{Err}^{\max}, \text{LocalErr}^{\max}/\text{Err}^{\max}\} < \delta_{\text{tol}}$ **then**
    $k_n \leftarrow \tilde{k}$
    break

```
%If the search ''fails'', simply select k within the range [l, u] that has the minimal error
```

if $k_n = k_n^{\max} + 1$ **then**
  $k_n \leftarrow \operatorname{argmin} \mathbf{Err}(l : u)$

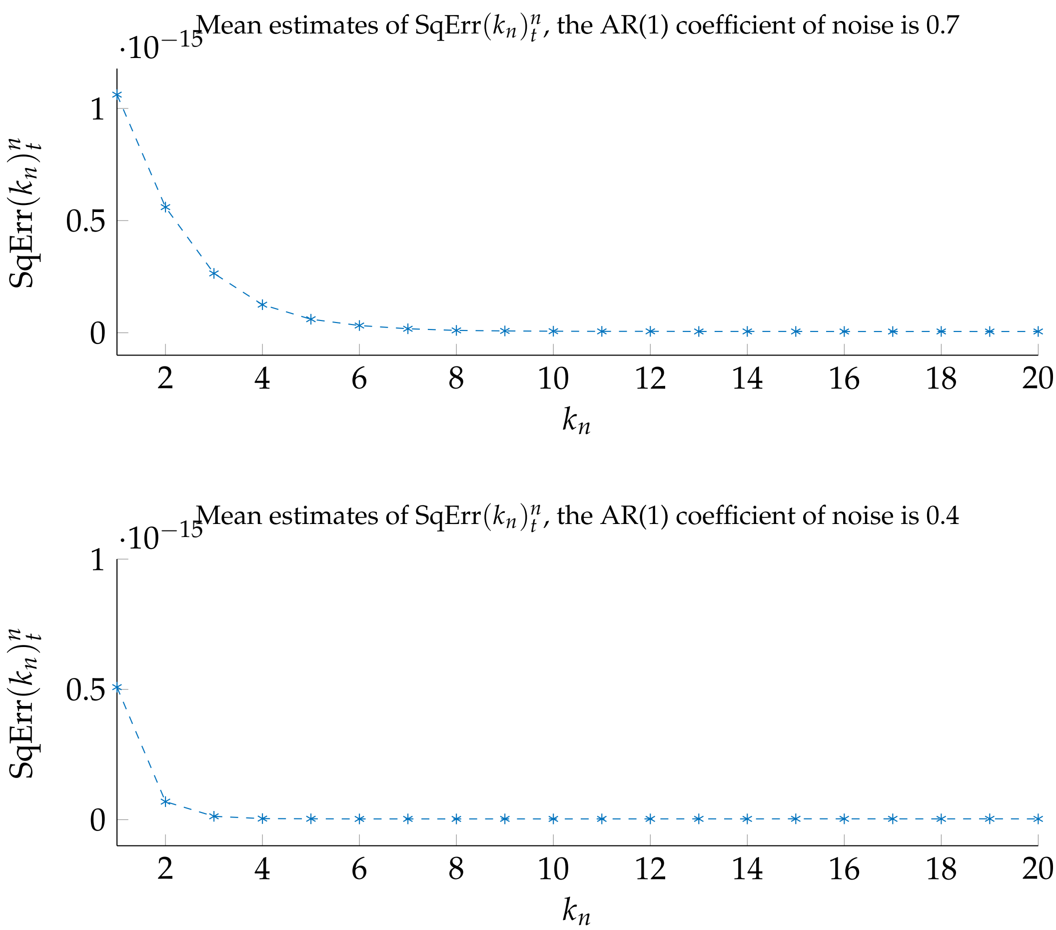

Figure F.1: Mean estimates of $\mathrm{SqErr}(k_n)_t^n$ (defined in (F.2)) based on 1,000 simulations. The AR(1) coefficients of the noise process are set to 0.7 and 0.4 to get the top and bottom panel.

F.1.1 The algorithm to choose k_n

Algorithm 1 is designed to select the tuning parameters k_n, which control the lengths of the non-overlapping intervals in the ReMeDI estimators, in which we utilize the error function (F.2).

F.2 Methodology 2

In this section, we consider a parametric model and study the "optimal" k_n, which minimizes the mean squared error (MSE).

Let n be the number of observations, and let the efficient price satisfy

$$X_{i+1}^n = X_i^n + \frac{1}{\sqrt{n}}\eta_i, \quad \varepsilon_{i+1} = \rho\varepsilon_i + u_i, \quad Y_i^n = X_i^n + \varepsilon_i.$$

where $\{\eta_i\}_i, \{u_i\}_i$ are normal i.i.d. with mean zero and variance $\sigma_\eta^2, \sigma_u^2$, respectively.

Given $\ell \in \mathbb{N}^*$, we have

$$\mathbb{E}\Big((Y_i^n - Y_{i-k_n}^n)(Y_{i+\ell}^n - Y_{i+\ell+k_n}^n)\Big) = r_\ell - 2r(\ell + k_n) + r(\ell + 2k_n),$$

where $r(\cdot)$ is the autocovariance function of ε. Hence, the leading term of the bias term is given by $2\sigma_u^2\rho^{\ell+k_n}/(1-\rho^2)$ and the leading term of the variance is given by $2k_n^3\sigma_\eta^4/3n^2$; we can select k_n to minimize $\frac{4\sigma_u^4\rho^{2(\ell+k_n)}}{(1-\rho^2)^2} + \frac{2k_n^3\sigma_\eta^4}{3n^2}$. Therefore, the optimal k_n^* should satisfy

$$(k_n^*)^2\rho^{-2k_n^*} = a(\rho, \sigma_u^2/\sigma_\eta^2)n^2, \text{ where } a(\rho, \sigma_u^2/\sigma_\eta^2) := \frac{8(\sigma_u^2/\sigma_\eta^2)^2 \log\left(\frac{1}{\rho}\right)\rho^{2\ell}}{(1-\rho^2)^2}. \tag{F.3}$$

Several observations are immediate: (1) the function $k \mapsto k^2\rho^{-2k}$ is increasing; (2) the function $a(\cdot)$ is increasing with respect to both variables. Hence, we conclude that there is unique k_n^* that satisfy the equality in (F.3), and k_n^* becomes larger when the dependence is stronger (larger ρ) or the noise-to-signal ratio is larger (larger σ_u^2/σ_η^2). In practice, the implementation of selecting k_n^* by this minimal MSE rule requires some proxies of $\rho, \sigma_u^2/\sigma_\eta^2$, which can be obtained by the maximum likelihood approach.

The two methodologies presented here are mostly useful as a guidance to select the tuning parameters if the sample size is small to moderate. Our experience with high-frequency data is that the ReMeDI approach is quite robust to the choices of k_n when applied to high-frequency data, see also the simulation studies presented in Section E.3. Hence, we stick to a fixed k_n in both the simulation and the empirical studies in the main text.

G Additional Empirical Studies

We replicate the empirical analysis presented in Li and Linton (2020) using the transaction prices of General Electric (GE) in January 2018. The estimates are presented in Figure G.1. Let's briefly highlight the findings: (1) noise tends to be positively autocorrelated, and the autocorrelation function remains positive up to 12 lags. This differs from our findings in the transaction prices of KO, where noise has more complicated autocorrelation patterns. (2) We observe larger discrepancy between the LA estimates and the ReMeDI estimates when estimations are performed on the entire sample of the month. As we analysed in Li and Linton (2020), the volatility burst of the efficient price (in a certain period of the month) is at least partially responsible for the increase in the finite sample bias of the LA estimators.

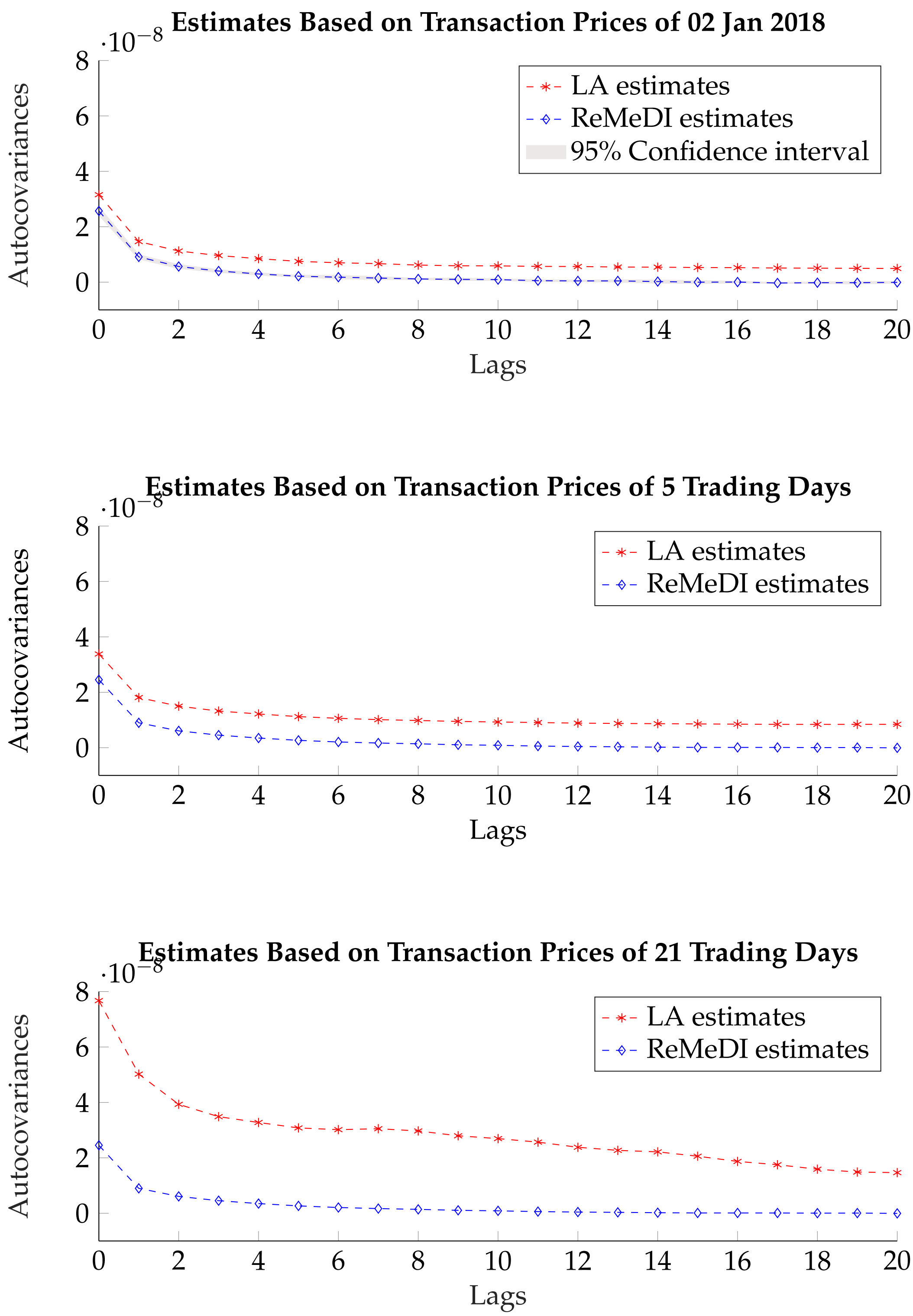

Figure G.1: Estimation of autocovariances of noise for GE in January 2018. In the top panel, we use the transaction prices of GE on 2 January 2018; in the middle panel, we use the transaction prices of GE in the second trading week (8 January 2018 to 12 January 2018); we employ the entire transaction prices of GE in January 2018 in the bottom panel. The tuning parameters for ReMeDI and LA are 10 and 6, respectively. The shaded area in the top panel represents the 95% confidence interval, and we set $i_n = 5, \phi_n = k_n^{3/5}/n$ to compute the asymptotic variances of the ReMeDI estimators, where n is the number of observations.

Next, we estimate the autocovariances of noise using the transaction prices of Citigroup (Citi) in January 2011. This is exactly the transaction prices used by Jacod et al. (2017) in their empirical analysis. Figure G.2 presents the estimates for each trading day in January 2011. The two estimates are very close, and the LA estimates are slightly larger. This is not surprised in view of the analysis by Li et al. (2020): the stock is extremely liquid (more than 10 observations per second on average) and the noise-to-signal ratio is also very large. It is in line with our analysis that the LA method works when both the data frequency and noise-to-signal ratio are high.

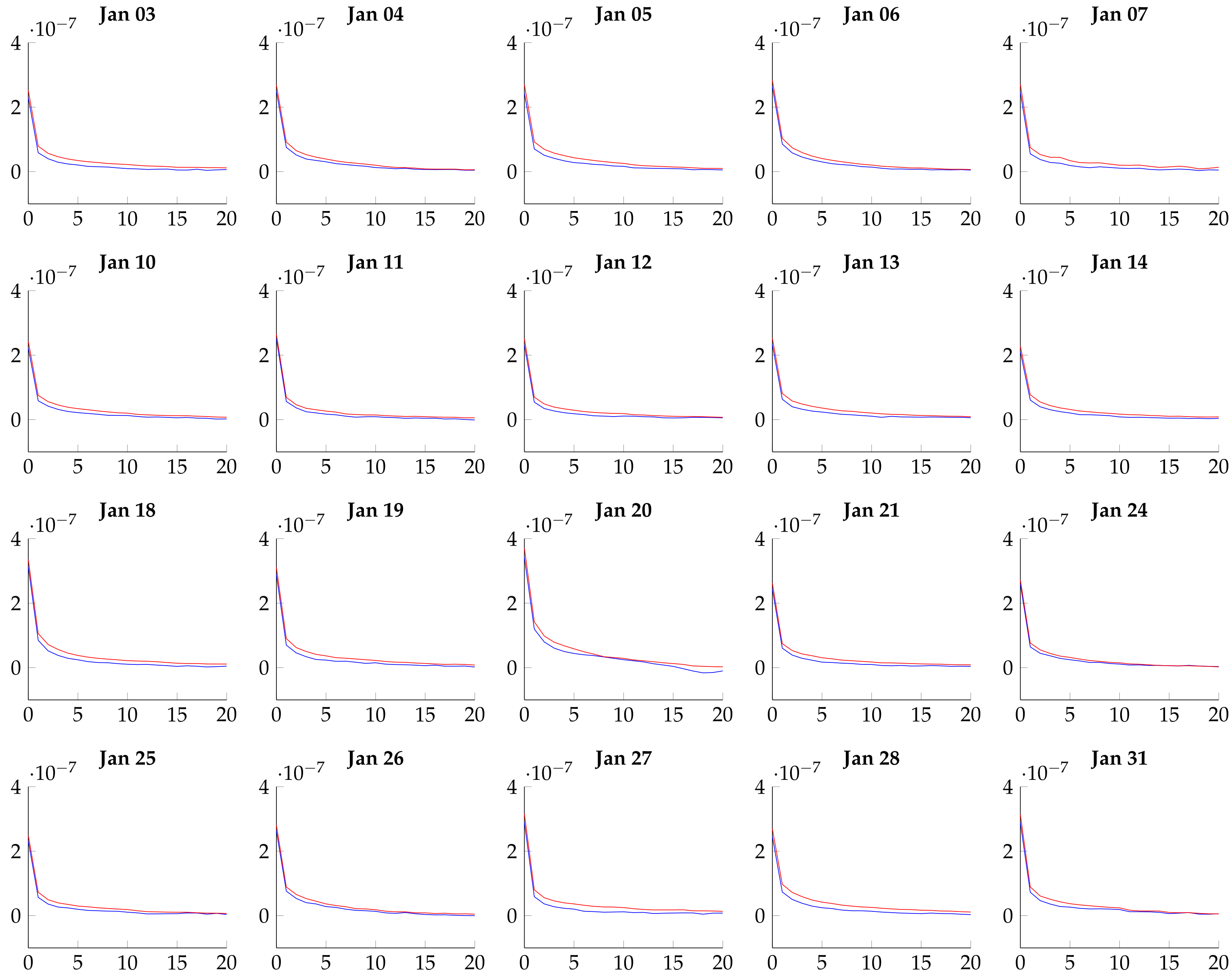

Figure G.2: ReMeDI (blue plots) and LA (red plots) estimates of autocovariances of noise for Citi in January 2011. The tuning parameters for ReMeDI and LA are 10 and 6, respectively.

References

JACOD, J., Y. LI, AND X. ZHENG (2017): "Statistical properties of microstructure noise," *Econometrica*, 85, 1133–1174.

——— (2019): "Estimating the integrated volatility with tick observations," *Journal of Econometrics*, 208, 80–100.

JACOD, J. AND M. ROSENBAUM (2013): "Quarticity and other functionals of volatility: efficient estimation," *Annals of Statistics*, 41, 1462–1484.

JACOD, J. AND A. N. SHIRYAEV (2003): *Limit Theorems for Stochastic Processes*, vol. 288, Springer-Verlag Berlin.

LI, Z. M., R. J. LAEVEN, AND M. H. VELLEKOOP (2020): "Dependent microstructure noise and integrated volatility estimation from high-frequency data," *Journal of Econometrics*, 215, 536–558.

LI, Z. M. AND O. LINTON (2020): "A ReMeDI for microstructure noise," https://papers.ssrn.com/abstract=3688778.