Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 11(8): 24733–24769.
DOI: 10.3934/math.2026996
Received: 12 April 2026
Revised: 18 July 2026
Accepted: 23 July 2026
Published: 12 August 2026

Research article

Estimation of integrated volatility in the presence of dependent and endogenous microstructure noise

Xin Luo[1,2,*]

[1] School of Economics and Management, Hubei University of Technology, Wuhan, China
[2] Hubei Circular Economy Development Research Center, Hubei University of Technology, Wuhan, China

* **Correspondence:** Email: luoxin@hbut.edu.cn.

Abstract: In this paper, a new econometric procedure is proposed to estimate the integrated volatility (IV) of the efficient price process in the presence of general dependent and endogenous market microstructure noise. A consistent estimator for IV, referred to as pre-averaging kernel realized volatility (PKRV$(Y)^{(n)}$), is first developed, based on a combination of the pre-averaging method and the realized kernel technique. The stable convergence of the proposed estimator is proved with a rate of $n^{-1/6}$. Next, to improve the performance of PKRV$(Y)^{(n)}$ in the presence of jumps, the threshold pre-averaging kernel realized volatility (TPKRV$(Z)^{(n)}$) is further proposed as an estimator for IV under market microstructure noise and jumps. Furthermore, the asymptotic properties of PKRV$(Y)^{(n)}$ and TPKRV$(Z)^{(n)}$, such as the consistency and central limit theorems, are established. A scheme for determining the optimal window width, the only parameter in PKRV$(Y)^{(n)}$ and TPKRV$(Z)^{(n)}$, is also proposed. Extensive simulation studies demonstrate the excellent performance of the proposed estimators, while empirical studies further illustrate their validity and accuracy.

Keywords: high-frequency; integrated volatility; market microstructure noise; dependent and endogenous noise; pre-averaging method
Mathematics Subject Classification: 62F12, 62M10, 62M20, 62P05

1. Introduction

The volatility of asset prices is critical to economic and financial decisions, risk management, asset pricing, and asset allocation. Andersen and Bollerslev [1] proposed the sum of squared intraday returns from high-frequency data as an estimator for the integrated volatility (IV) of the latent efficient price process, referred to as realized variance (RV), which soon became a popular estimator for IV.

If such high-frequency data were not contaminated by market microstructure noise, the RV of the observed transaction price* process following an Itô semimartingale process [1, 2] would theoretically be a consistent estimator for IV. When estimating the IV by RV at a high sampling frequency, we will be forced to face an "infinity" disaster from the accumulation of market microstructure noise, which can be attributed to various market microstructure effects including, but not limited to, bid-ask spreads, information asymmetries, the discreteness of the data, transaction costs, and rounding errors.

Market frictions make markets inefficient and cause the observed price to deviate from the efficient price (or "frictionless equilibrium price"), meaning that the observed price Y can be decomposed into the unobservable efficient price X plus a noise component ε, $Y_t = X_t + \varepsilon_t$. Accurate volatility estimation, in turn, provides the foundation for subsequent forecasting and empirical investigations across a wide range of market settings. These include both methodological developments in forecasting frameworks [3–5] and empirical analyses of macroeconomic and cross-market drivers of volatility [6, 7].

Estimating the IV of the efficient price process X from high-frequency data is an emblematic topic. The noise component appears only when transaction occurs, so the noise component can be explained as a measurement error. It is the IV of the efficient price process that describes the volatility of the asset price process instead of the RV of the observed price process with market microstructure noise. Therefore, how to eliminate the effect of market microstructure noise on estimation of IV is one of the core issues in high-frequency financial econometrics.

In general, the efficient price process is assumed to follow an Itô semimartingale process. However, the microstructure noise process has been theoretically and empirically proved to be serially dependent, nonstationary, and correlated with the efficient returns [8–10]. A common assumption in market microstructure is that the noise is independently and identically distributed (i.i.d.). Under this assumption, several IV estimators have been proposed to account for the contamination of the efficient price caused by microstructure noise.

The two-scale realized volatility (TSRV) estimator proposed by [11] provides a consistent estimator of IV with the convergence rate $n^{-1/6}$. Zhang [12] developed the multi scale realized volatility (MSRV) with the optimal convergence rate $n^{-1/4}$. The simple realized kernel estimator of IV was proposed by [13], in which the first-order autocorrelation of the observed price process was used to eliminate the bias of RV. However, the simple realized kernel estimator is not a consistent estimator for IV. Hansen and Lunde [8] proposed the realized kernel estimator, which uses high-order autocovariances of the observed price process and yields an unbiased but inconsistent estimate of IV. To improve upon this estimator, Barndorff-Nielsen et al. [14] proposed the flat top realized kernels, which are consistent estimators for IV and achieve the optimal rate $n^{-1/4}$. Aït-Sahalia et al. [15] proposed the maximum likelihood estimation (MLE) approach under constant volatility. This approach was subsequently extended by [16–18]. The econometric literature on pre-averaging methods was initiated by Podolskij and Vetter [19], who proposed a consistent estimator for IV, termed the modulated realized volatility (MRV). Subsequently, Jacod et al. [20, 21] developed the pre-averaging approach further by adopting moving-window weighted averaging, which substantially reduces the conditional variance of the estimator and has since become one of the standard estimation methods for IV.

We know that the i.i.d. noise assumption is contrary to empirical evidence contained in [8, 22, 23], among others. To take into account the effect of the autocorrelation of market microstructure noise on the estimation of IV, Aït-Sahalia et al. [24] developed the extended TSRV estimator, which is robust

*In this paper, "price" always refers to the "logarithmic price".

to serially dependent microstructure noise. Da and Xiu [17] extended the likelihood-based method proposed by [15, 16] to allow for serially dependent noise. Li and Linton [25] proposed a two-step approach for estimating IV, wherein they relaxed the i.i.d. noise assumption and assumed that the noise process is stationary and independent of the efficient price process and that its autocovariance decays exponentially.

Indeed, the assumptions about microstructure noise in these studies are still not consistent with the empirical evidence documented by [8], who showed that microstructure noise is negatively correlated with efficient returns; that is, the microstructure noise behaves endogenously. This endogeneity is theoretically justified in [9] and empirically supported by [10, 26]. Estimators that ignore the endogeneity of microstructure noise are likely to underestimate or overestimate IV. Jacod et al. [27] studied IV estimation with tick observations and allowed the noise to depend on the efficient price process X. However, despite this dependency, the noise and the efficient returns remain uncorrelated in their setup. Andersen et al. [10] examined the impact of endogenous microstructure noise on IV estimation but did not propose estimators for IV. Consequently, the assumptions on the microstructure noise process in the existing literature remain overly restrictive, leading to biased estimators for IV.

Market microstructure noise in high-frequency financial data has been empirically found to be serially dependent and endogenous [8, 10]. Ignoring these features induces substantial IV estimation bias, with adverse implications for risk management and option pricing. Yet the prevailing pre-averaging and kernel-based methods, designed under assumptions of i.i.d. noise or noise that is uncorrelated with efficient returns, fail to capture the complexity of real-world noise. This motivates the search for a more robust estimator.

Specifically, in [13, 28], the simple kernel estimator, which uses high-order autocovariances to bias-correct RV, has been shown to be unbiased but not consistent under finite-dependent noise. The pre-averaging strategy can smooth out uncorrelated noise fluctuations, thereby reducing the variance. This raises the question: Can the pre-averaging strategy be systematically integrated into the kernel-based estimator to develop an enhanced estimator for IV? The present paper is motivated by this question.

In this paper, we relax the i.i.d. noise assumption and obtain an unbiased and consistent estimator of IV that allows the market microstructure noise to be serially dependent and endogenous. We assume only that the noise process is difference-stationary and has a finite correlation time, which is consistent with financial market evidence [8, 9, 24]. Apart from introducing an optimal window width, we impose no parametric restrictions on the noise's distribution.

Several pre-averaging estimators have been proposed recently, but none addresses endogeneity. Liu et al. [29] extended the pre-averaging method to handle multiple observations per timestamp, a challenging data feature, while retaining the classical i.i.d. noise assumption. In a recent contribution, [30] proposed a pre-averaging estimator for IV that builds upon their earlier realized moments of disjoint increments (ReMeDI) methodology developed in [25], and is robust to serially dependent microstructure noise. However, both approaches implicitly maintain the assumption that the noise and the efficient returns are uncorrelated, which rules out endogenous noise, contrary to empirical facts [8, 10].

The proposed pre-averaging kernel realized volatility ($\mathrm{PKRV}(Y)^{(n)}$) estimator differs fundamentally from these pre-averaging methods. First, whereas [30] can accommodate serially dependent noise but cannot handle endogeneity, $\mathrm{PKRV}(Y)^{(n)}$ accommodates both. Second, pre-averaging methods such as [30] replace raw returns with pre-averaged returns in RV, which yields a biased estimator that requires bias correction. In contrast, $\mathrm{PKRV}(Y)^{(n)}$ applies pre-averaged returns to the simple realized kernel of [13], producing an unbiased and consistent estimator by construction even when noise is serially dependent and endogenous. Third, we introduce a data-driven correlation time test for window selection, a tool not available in [30]. Consequently, the two approaches are complementary: Their method is suitable for settings where noise is dependent but uncorrelated with the efficient returns, whereas $\mathrm{PKRV}(Y)^{(n)}$ is designed for the empirically relevant scenario where noise is both dependent and endogenous.

This paper makes four contributions. First, we propose a new estimator, $\mathrm{PKRV}(Y)^{(n)}$, which combines pre-averaging with a simple kernel correction [13] to eliminate bias arising from serially dependent and endogenous noise. Second, we prove its consistency and asymptotic normality under the sole assumptions that the noise is difference-stationary and has a finite correlation time. Third, we develop a data-driven correlation time test to select the window's width, directly bridging estimation and inference. Fourth, we extend the estimator to handle jumps via its thresholding version, TPKRV. Simulation and empirical studies demonstrate excellent finite-sample performance, notably the absence of negative IV estimates at the selected optimal window width.

The remainder of this paper is organized as follows. In Section 2, we state the basic notation, definition, and assumptions of the efficient price process and market microstructure noise. In Section 3, we give an asymptotic theory of our proposed estimator $\mathrm{PKRV}(Y)^{(n)}$ for IV and further develop the estimator $\mathrm{TPKRV}(Z)^{(n)}$ in the presence of market microstructure noise and jumps on the basis of the threshold technique. Section 3 also details the scheme for determining the optimal window width. Section 4 reports extensive simulation studies. Empirical studies are presented in Section 5. Section 6 concludes the paper. All proofs are given in the Appendix.

2. Definition, notation, and background

The efficient market hypothesis says that the efficient price (x) must follow a semimartingale. The most elegant semimartingale price process is a Brownian semimartingale defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$

$$dX_t = \mu_t dt + \sigma_t dw_t, \tag{2.1}$$

where $w = (w_t)$ is a standard Wiener process, $\mu = (\mu_t)$ is a locally bounded predictable drift process, and $\sigma = (\sigma_t)$ is a *càdlàg* volatility process—all adapted to some common filtration $\mathcal{F}_t$. Here, σ_t is a random function, and σ_t^2 is Lipschitz (almost surely). The significant measure is the IV over one period (usually one day)

$$IV = \int_0^1 \sigma_t^2 \, dt.$$

A natural way to estimate this object is to use the quadratic variation

$$[X, X]_T = \sum_{t_i} (X_{t_{i+1}} - X_{t_i})^2, \tag{2.2}$$

where the values of X_{t_i} are all of the observed price processes in [0,T], and $0 = t_0 < t_1 < \cdots < t_n = T$. The estimator $\sum_{t_i}(X_{t_{i+1}} - X_{t_i})^2$ is commonly used and generally called "realized variance" [1]. Assuming the price series follows a Brownian process, the asymptotic behavior of the realized variance is justified by the theoretical results in stochastic processes stating that

$$\text{plim} \sum_{t_i} (X_{t_{i+1}} - X_{t_i})^2 = \int_0^1 \sigma_t^2 \, dt. \tag{2.3}$$

In the following stochastic analysis on the estimators of IV, we condition on the volatility path σ_t^2. Thus σ_t^2 can be treated as deterministic even though we view the volatility path as random.

However, in financial markets, many market frictions, such as transaction costs, price discreteness, inventory holdings, information asymmetries, and measurement errors, may cause the observed prices to deviate from the efficient prices. So, we can express the observed asset price Y as the sum of the market microstructure noise ε and the efficient price X as follows:

$$Y_{t_i} = X_{t_i} + \varepsilon_{t_i}. \tag{2.4}$$

We construct intraday returns from equidistant observed price data under calendar time sampling (CTS). We partition the time interval $T = 1$ (one trading day) into n intervals of equal length, $\Delta_n = 1/n$. The price Y observed at the time points $t_i = i\Delta_n$, $i = 0, \ldots, n$ is denoted $Y_i = Y_{t_i}$. The raw observed price data are irregularly spaced. The observed prices under CTS must be artificially constructed from the raw price data. Here, we adopt the previous tick method [31] to get the observed price data under CTS. Given the irregularly spaced observed transaction price data at the time points $t_0 < \cdots < t_n$, one can construct a price at time $\tau \in [t_j, t_{j+1})$ by using $Y_\tau \equiv Y_{t_j}$.

We acknowledge that this transformation discards some information: Multiple transactions within a given interval are replaced by a single price, and trade by trade details such as the order flow and volume are lost. Nevertheless, this loss is acceptable in our framework. First, the pre-averaging step aggregates returns over a window of length k_n, which smooths out uncorrelated noise fluctuations and reduces conditional variance at a rate $O(\frac{1}{k_n})$. Consequently, the mild information loss from CTS does not affect the consistency of our estimator. Second, CTS aligns the observed data with the continuous-time semimartingale model, providing a tractable framework for deriving asymptotic properties such as the central limit theorem. Thus, while CTS is imperfect, it represents a reasonable trade-off for achieving reliable volatility estimation.

The X and ε are latent but their sum Y is observed. The realized variance at a frequency n is defined by

$$RV^{(n)} = \sum_{i=1}^{n} (\Delta Y_{t_i})^2 = \sum_{i=1}^{n} (\Delta X_{t_i})^2 + 2\sum_{i=1}^{n} \Delta X_{t_i} \, \Delta\varepsilon_{t_i} + \sum_{i=1}^{n} (\Delta\varepsilon_{t_i})^2. \tag{2.5}$$

Almost all the existing literature assumes that ε is an i.i.d. process and is independent of the efficient price process X, which implies that

$$\sum_{i=1}^{n} (\Delta Y_{t_i})^2 = \sum_{i=1}^{n} (\Delta X_{t_i})^2 + \sum_{i=1}^{n} (\Delta\varepsilon_{t_i})^2, \tag{2.6}$$

under the i.i.d. noise assumption. When the sampling frequency increases to infinity, the second term on the right of Eq (2.6) diverges to infinity. Thus, the RV does not estimate the true IV, and the IV is submerged in the sea of noise. Therefore, denoising is a key issue in volatility estimation from high-frequency data. Eq (2.6) shows that the i.i.d. noise assumption simplifies the estimation of IV. However, this assumption has been rejected by empirical evidence [8, 22, 24].

The recently proposed approach for estimating IV [32] relaxes the i.i.d. noise assumption and allows for dependent noise, but it still assumes that the noise is independent of the efficient price. As a result, it underestimates the IV. In contrast, we will relax the noise assumption to better reflect empirical facts and aims to obtain an unbiased and consistent estimator of IV in the presence of dependent and endogenous microstructure noise. We make the following assumptions on the microstructure noise process, ε.

Assumption 2.1 (The difference-stationary assumption)**.** The noise increment process, $\Delta\varepsilon_i$, is covariance-stationary with mean zero, such that its autocovariance function is defined by

$$\pi^{(h)} = \mathbb{E}[\Delta\varepsilon_i \, \Delta\varepsilon_{i+h}],$$

and its covariance with efficient price increments, ΔX_i, is defined by

$$\gamma^{(h)} = E[\Delta X_i \, \Delta\varepsilon_{i+h}].$$

The i.i.d. noise assumption extensively adopted in the existing literature trivially satisfies the difference-stationary assumption, which has $\pi^{(1)} = -E(\varepsilon_i^2) = -\omega^2$ and $\pi^{(h)} = 0$ for all $h > 1$, and $\gamma^{(h)} = 0$ for all $h \neq 0$, and the Ornstein-Uhlenbeck serially dependent noise process is another simple example that satisfies Assumption 2.1.

Assumption 2.2 (The finite correlation time assumption)**.** The noise process has finite dependence and finite endogeneity in the sense that $\pi(s) = E(\varepsilon(t)\,\varepsilon(t+s)) = 0$ for all $s > \theta_0$ for some finite $\theta_0 \geq 0$, and $E[\varepsilon(t) \mid X(s)] = 0$ for all $|t - s| \geq \theta_0$ for some finite $\theta_0 \geq 0$.

The mixing condition on the market microstructure noise adopted in the recent literature [23, 27] satisfies Assumption 2.2. In tick time, $n\theta_0 \to \infty$ $(n \to \infty)$ implies the mixing condition.

The difference-stationary assumption does not require the microstructure noise to be stationary, and also accommodates possible diurnal features of the noise so that the size of the noise may change over time. The finite correlation assumption is compatible with the empirical evidence, and is also a natural assumption. In sum, Assumptions 2.1 and 2.2 are rather weak and completely compatible with empirical evidence in the existing literature.

We are now ready to characterize the parameters $\pi^{(h)}$, $\gamma^{(h)}$ under CTS.

Lemma 2.1. Assume that Assumptions 2.1 and 2.2 hold, and let $k_n \geq \lceil n\theta_0 \rceil + 1$. Then

$$\sum_{h=-k_n}^{k_n} \gamma^{(h)} = 0, \tag{2.7}$$

$$\sum_{h=-k_n}^{k_n} \pi^{(h)} = 0. \tag{2.8}$$

The two equalities indicate that, over a window wider than the noise correlation time, the cumulative covariance between the efficient price increments and noise increments and the cumulative autocovariance of noise increments sum to zero. Economically, this reflects a form of self-canceling short memory in microstructure frictions: Although the noise may be serially dependent and endogenous, its influence does not persist beyond a certain horizon. This property is of critical importance, as it enables the proposed estimator to eliminate noise-induced bias without imposing i.i.d. noise or independence of the noise from the efficient price. In practice, it justifies using a finite window to average out the impact of frictions such as bid–ask spreads, rounding errors, and information asymmetry, all of which generate only short-lived distortions in high-frequency prices.

This design of $\mathrm{PKRV}(Y)^{(n)}$ effectively mitigates the bias and conditional variance induced by dependent and endogenous microstructure noise. Intuitively, this efficacy stems from two principal mechanisms.

(1) Asymptotic bias elimination via covariance construction: The bias terms in $\mathrm{PKRV}(Y)^{(n)}$ associated with noise take the form $\gamma^{(h)}$ and $\pi^{(h)}$. The design of $\mathrm{PKRV}(Y)^{(n)}$ yields the sum of bias terms as

$$\sum_{h=-k_n}^{k_n} (\gamma^{(h)} + \pi^{(h)}).$$

By Lemma 2.1, it follows that

$$\lim_{k_n\to\infty} \sum_{h=-k_n}^{k_n} (\gamma^{(h)} + \pi^{(h)}) = 0,$$

thereby eliminating the asymptotic bias.

(2) Variance reduction via bandwidth averaging: Returns are averaged over a bandwidth of k_n periods, yielding repeated quasi-independent measurements of the same latent price. This averaging smooths out uncorrelated noise fluctuations, reducing the conditional variance at the rate $O(\frac{1}{k_n})$.

3. Asymptotic theory

3.1. *Asymptotic theory of* $\mathrm{PKRV}(Y)^{(n)}$

For any process $V = (V_t)_{t\geq 0}$, we define the average increments in the process V as

$$\Delta V_i = V_i - V_{i-1}, \tag{3.1}$$

$$\overline{V}_i = \frac{1}{\sqrt{k_n}} \sum_{j=1}^{k_n} \Delta V_{i+j}. \tag{3.2}$$

Recall that in our notation, we do not observe the efficient process X but the observed price Y. So the estimator of IV should be constructed on the basis of the values Y_i only, and we propose the following estimator for IV based on the average returns $\overline{Y}_i$:

$$\mathrm{PKRV}(Y)^{(n)} = \sum_{i=k_n}^{n-2k_n} \overline{Y}_i^2 + \sum_{i=k_n}^{n-2k_n} \overline{Y}_i\,\overline{Y}_{i+k_n} + \sum_{i=k_n}^{n-2k_n} \overline{Y}_i\,\overline{Y}_{i-k_n}. \tag{3.3}$$

The second term and the third term on the right-hand side of Eq (3.3) are intended to offset the variance and covariance of the market microstructure noise.

Theorem 3.1. Assume that

$$\int_0^T \sigma_t^4\, dt < \infty,$$

$$E|\varepsilon_i|^{4+\eta} < \infty,$$

and

$$E(X_i\varepsilon_{i+h})^{2+\eta} < \infty$$

for some $\eta > 0$ and any integer h. If Assumptions 2.1 and 2.2 hold, also let $k_n \to \infty, n \to \infty$ and $k_n/n \to 0$. We then have,

$$E(\mathrm{PKRV}(Y)^{(n)} \mid \{\sigma_t^2\}_0^1) = \int_0^1 \sigma_t^2\, dt + \mathrm{O}\left(\frac{k_n}{n}\right), \tag{3.4}$$

holds, and when $k_n = k_n^* = cn^{2/3}$, the asymptotic behavior is given by

$$\frac{n^{1/6}\left(\mathrm{PKRV}^{(n)} - \int_0^1 \sigma_t^2\, dt\right)}{\sqrt{\Gamma}} \xrightarrow{\mathcal{L}} N(0,1), \tag{3.5}$$

where Γ is the asymptotic variance of the estimator $PKRV(Y)^{(n)}$.

Remark 3.1. *It is worth noting that the proposed estimator* $\mathrm{PKRV}(Y)^{(n)}$ *achieves the convergence rate* $n^{-1/6}$*, which is slower than the optimal convergence rate* $n^{-1/4}$ *attainable under i.i.d. noise [12, 14] or under specially structured dependent noise with exponentially decaying autocovariance [32]. However, these faster rates rely on much stronger assumptions. In contrast, our Assumption 2.2 only requires that the noise has a finite correlation time in calendar time, which is substantially weaker and more realistic for financial data. The slower convergence rate* $n^{-1/6}$ *is the cost we bear for considering dependent and endogenous noise. The simulation's results confirm that despite the slower asymptotic convergence rate, the finite-sample performance of the proposed estimator exhibits smaller bias and standard deviations than the alternative estimator under dependent and endogenous noise scenarios, and negative estimates essentially vanish at the optimal window width. We believe this rate-robustness trade-off is fully justified, given the empirical reality of dependent and endogenous microstructure noise.*

Remark 3.2. *In the financial markets, especially in emerging markets, serious deviations from* $k_n/n \to 0$ *could occur in finite samples. The subsequent empirical analyses will confirm this viewpoint. In general, an estimator derived from asymptotic results can, however, behave very differently in finite samples. With the estimator* $\mathrm{PKRV}(Y)^{(n)}$*, the finite sample bias results from the fact that the first* k_n *and the last* $2k_n$ *terms, as seen in Eq (3.3), have been discarded. Therefore, we propose a finite-sample bias-corrected estimator*

$$\mathrm{PKRV}(Y)^{(n)(adj)} = \frac{n}{n-3k_n}\ \mathrm{PKRV}(Y)^{(n)}, \tag{3.6}$$

which is trivially an unbiased and consistent estimator of IV.

The unbiasedness result of the estimator $\mathrm{PKRV}(Y)^{(n)}$ *shows that even in the presence of dependent and endogenous noise, it correctly targets the true IV, with only a finite-sample bias that vanishes as the ratio of window width to sample size tends to zero. From an economic perspective, the estimator recovers the asset's fundamental volatility, which is a key input for option pricing, value at risk calculations, and portfolio allocation, without systematic distortion from market frictions. This robustness to more realistic noise features comes at the cost of a slower convergence rate, but the finite-sample bias correction prevents systematic over- or underestimation of volatility, even in small samples, an essential feature for real-time risk monitoring using daily estimates.*

3.2. *Asymptotic theory of* TPKRV$(Z)^{(n)}$

In this section, we explore the effect of jumps on the PKRV$(Y)^{(n)}$ to get an estimator of IV when both microstructure noise and jumps are present. For this purpose, we consider the model

$$Z_t = Y_t + J_t, \tag{3.7}$$

where $Y = (Y_t)$ is a noisy diffusion process defined by Eq (2.4) and $J = (J_t)$ denotes a finite-activity jump process; that is, J exhibits finitely many jumps on compact intervals. A typical example of a finite-activity jump process is the compound Poisson process. Note that $J_t = \sum_{i=1}^{N_t} k_{\tau_i}$, where $N = (N_t)$ is a Poisson process with an intensity λ, and k_{τ_i} denotes the jump size at the jump location τ_i. The jump sizes k_{τ_i} are i.i.d., and independent of the jump process N. We also assume that (N_t) is independent of the efficient price process X and the microstructure noise process ε. The pre-average of the observed price increments can be expressed as

$$\overline{Z}_i = \overline{Y}_i + \overline{J}_i, \tag{3.8}$$

Lemma 3.1. If we assume that $J \perp X$ and $J \perp \varepsilon$, and J is a compound Poisson process with finite activity of jumps, then

$$E\Big(\sum_{i=k_n}^{n-2k_n} \overline{Z}_i^2\Big) = E\Big(\sum_{i=k_n}^{n-2k_n} \overline{Y}_i^2\Big) + E\Big(\sum_{i=k_n}^{n-2k_n} k_{\tau_i}^2\Big), \tag{3.9}$$

and

$$E(\mathrm{PKRV}(Z)^{(n)}) = E(\mathrm{PKRV}(Y)^{(n)}) + E\left(\sum_{i=k_n}^{n-2k_n} k_{\tau_i}^2\right), \tag{3.10}$$

where $\sum_{i=k_n}^{n-2k_n} k_{\tau_i}^2$ is the variance of the jump process, and this term comes only from the term $\sum_{i=k_n}^{n-2k_n} \overline{Z}_i^2$. Therefore, we can use the threshold technique to eliminate the effects of jumps on estimate of IV. In the framework of the threshold technique [33, 34], we propose the following threshold pre-averaging kernel ralized volatility:

$$\mathrm{TPKRV}(Z)^{(n)} = \sum_{i=k_n}^{n-2k_n} \overline{Z}_i^2\, 1_{\{|\overline{Z}_i| \le \Theta(\Delta_n)\}} + \sum_{i=k_n}^{n-2k_n} \overline{Z}_i\, \overline{Z}_{i+k_n} + \sum_{i=k_n}^{n-2k_n} \overline{Z}_i\, \overline{Z}_{i-k_n}, \tag{3.11}$$

where $\Theta(\Delta_n)$ is the threshold function.

Some simple variance calculations show that the pre-averaging of efficient price increments, microstructure noise increments, and jump increments are given as

$$\overline{X}_i = \frac{1}{\sqrt{k_n}} \sum_{j=1}^{k_n} (X_{i+j} - X_{i+j-1}) = O_p(\Delta_n^{1/2}),$$

$$\overline{\varepsilon}_i = \frac{1}{\sqrt{k_n}} \sum_{j=1}^{k_n} (\varepsilon_{i+j} - \varepsilon_{i+j-1}) = O_p(\Delta_n^{1/3}), \quad \text{when } k_n = Cn^{2/3},$$

$$\overline{J}_i = \frac{1}{\sqrt{k_n}} \sum_{j=1}^{k_n} (J_{i+j} - J_{i+j-1}).$$

After pre-averaging, the smoothed increments from efficient price and microstructure noise have the size $\Delta_n^{1/2}$ and $\Delta_n^{1/3}$, respectively. However, if there are jumps in the interval $[i, i+k_n]$, then $\overline{J}_i \neq 0$. The pre-averaging does not eliminate jumps. To identify the jumps in the pre-averaging observations $\overline{Z}_i$, the threshold function is required to satisfy the following assumption.

Assumption 3.1. The threshold function $\Theta(\Delta_n)$ is a deterministic function of the step length Δ_n such that

$$\lim_{\Delta_n \to 0} \Theta(\Delta_n) = 0, \tag{3.12}$$

$$\lim_{\Delta_n \to 0} \frac{\Delta_n^{1/3}}{\Theta(\Delta_n)} = 0. \tag{3.13}$$

Power functions $\Theta(\Delta_n) = C\Delta_n^{\alpha}$ for any $\alpha \in (0, 1/3)$ and $C \in R$ are possible choices. The threshold function $\Theta(\Delta_n)$ can be used to identify whether jumps occur in the interval $[t_i, t_{i+k_n}]$ or not.

Lemma 3.2. If Assumption 3.1 and the assumptions of the Lemma 3.1 are satisfied, we have

$$\Delta_n^{-\beta}\Big(\mathrm{TPKRV}(Z)^{(n)} - \mathrm{PKRV}(Y)^{(n)}\Big) \xrightarrow{p} 0, \tag{3.14}$$

where $\beta \in (0, 1/6)$.

If $\mathrm{TPKRV}(Z)^{(n)}$ is used directly as an estimator of IV in the presence of microstructure noise and jumps, its conditional variance and bias become substantial in finite samples due to too few jump observations. This leads to large measurement errors. Therefore, we propose a modified version of $\mathrm{TPKRV}(Z)^{(n)}$ to reduce these errors.

$$\mathrm{TPKRV}(Z)^{(n)} = \sum_{i=k_n}^{n-2k_n} \overline{Z}_i^2\, 1_{\{|\overline{Z}_i| \le \Theta(\Delta_n)\}} + \sum_{i=k_n}^{n-2k_n} \overline{Z}_i\, \overline{Z}_{i+k_n}\, 1_{\{|\overline{Z}_i| \le \Theta(\Delta_n)\}} + \sum_{i=k_n}^{n-2k_n} \overline{Z}_i\, \overline{Z}_{i-k_n}\, 1_{\{|\overline{Z}_i| \le \Theta(\Delta_n)\}}, \tag{3.15}$$

We can obtain the finite sample bias corrected estimator

$$\mathrm{TPKRV}(Z)^{(n)(adj)} = \frac{n}{n - 3k_n}\, \mathrm{TPKRV}(Z)^{(n)}. \tag{3.16}$$

Theorem 3.2. If the assumptions of Theorem 3.1 and Lemma 3.1 are satisfied, then we have

$$\Delta_n^{-\beta}\Big(\mathrm{TPKRV}(Z)^{(n)(adj)} - \int_0^1 \sigma_t^2\, dt\Big) \xrightarrow{p} 0, \tag{3.17}$$

where $\beta \in (0, 1/6)$. Note that the convergence rate of $\mathrm{TPKRV}(Z)^{(n)(adj)}$ is lowered to be less than $n^{-1/6}$ due to the threshold technique.

3.3. Determining of the optimal window width k_n^*

The optimal window width k_n^* must satisfy two criteria. First, it should be greater than the correlation time, i.e., $k_n^* \ge \lceil n\theta_0 \rceil + 1$. Second, it should minimize the mean square error (MSE) of the estimator. Under the minimum MSE criterion, we prove $k_n^* \sim cn^{2/3}$. To determine k_n^* in practice, we propose the following two-step procedure.

3.3.1. Determining of correlation time

Following the literature that connects volatility estimation to statistical inference [35], we develop a test for the noise correlation time to justify the window width choice in our $\mathrm{PKRV}(Y)^n$ estimator.

Here, we propose the statistic $t_{(k_n,k'_n)}$ to test the correlation time, following the framework of [36], with $d_t = \mathrm{PKRV}(Y)^{(n)}(k_n) - \mathrm{PKRV}(Y)^{(n)}(k'_n)$. Under the null hypothesis that both k_n and k'_n are not shorter than the correlation time $\lceil n\theta_0 \rceil + 1$, the random variables d_t are asymptotically uncorrelated because they capture only the estimation errors from the remaining uncorrelated noise. Under the null hypothesis, $E(d_t) = 0$. Then, by the central limit theorem, the t-statistic converges to a mixed normal distribution. Specifically, we construct the statistic $t_{(k_n,k'_n)}$ as follows:

$$t_{(k_n,k'_n)} = \frac{\sqrt{n}\,\overline{d}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2}}, \tag{3.18}$$

$$t_{(k_n,k'_n)} \xrightarrow{\mathcal{L}} N(0,1) \quad \text{as } n \to \infty. \tag{3.19}$$

We fix $k_n^{'}$ and set $k_n^{'} \geq \lceil n\theta_0 \rceil + 1$ in the hypothesis testing. When $k_n \geq \lceil n\theta_0 \rceil + 1$, $t_{(k_n,k'_n)}$ converges to standard normal distribution. When $k_n < \lceil n\theta_0 \rceil + 1$, the stochastic series (d_t) is no longer independent, and the statistic $t_{(k_n,k'_n)}$ does not converge to standard normal distribution. The null hypothesis takes the form

$$H_0 : k_n \geq \lceil n\theta_0 \rceil + 1. \tag{3.20}$$

We denote the alternative hypothesis

$$H_1 : k_n < \lceil n\theta_0 \rceil + 1. \tag{3.21}$$

In the empirical study, we take the following steps to conduct the hypotheses test.
(1) Setting a large enough $k_n^{'}$, e.g., 10 minutes or 15 minutes, which is believed to be larger than the correlation time.
(2) Setting a step Δs, e.g., $\Delta s = 30$ seconds, such that $k_n = k_n^{'} - \Delta s$.
(3) Testing whether the statistics $t_{(k_n,k'_n)}$ is 0 or not.
(4) If one accepts the null hypothesis, it means that $k_n \geq \lceil n\theta_0 \rceil + 1$. We continue to drop k_n as $k_n = k_n^{'} - \Delta s$, until the null hypothesis is rejected. The last k_n passing the $t_{(k_n,k'_n)}$ test is selected as the correlation time.
(5) To get a more accurate correlation time, one can decrease the $k_n^{'}$ and the step Δ_s, then repeat Step 2 to Step 4.

The statistic $t_{(k_n,k'_n)}$ is designed to detect the minimum correlation time beyond which the noise autocovariance and the cross-covariance between noise and efficient returns become negligible. In financial terms, this correlation time reflects how quickly the market microstructure frictions dissipate. For highly liquid stocks, the correlation time may be only a few seconds or minutes, whereas for less liquid assets or during volatile periods, it can be longer. By empirically determining this length from the data rather than imposing a fixed value, our procedure adapts to actual market conditions. This is economically important: Using a window width shorter than the correlation time leaves residual bias in the volatility estimate, while using an unnecessarily long window width inflates the conditional variance. The proposed test thus achieves a bias-variance trade-off tailored to the asset's microstructure environment, which is particularly valuable for high-frequency trading firms and regulators monitoring real-time volatility.

3.3.2. Determining the criterion of the optimal window width k_n^*

The correlation time determined by the $t_{(k_n,k_n')}$ test statistic can be seen as a lower bound for the optimal window width k_n^*. In empirical applications, one can select the optimal window width k_n^* by minimizing the MSE of the estimator over the range exceeding this lower bound. To make this choice more reliable, we seek as much evidence as possible. In particular, if the correct k_n^* is chosen, the probability of obtaining a negative IV estimate vanishes as the sampling frequency tends to infinity when one applies $\mathrm{PKRV}(Y)^{(n)(adj)}$ and $\mathrm{TPKRV}(Z)^{(n)(adj)}$. Consequently, this probability can also serve as an additional criterion for selecting k_n^*. Therefore, we define the optimal window width as the value that exceeds the correlation time and simultaneously minimizes the MSE and the probability of a negative estimate of IV.

4. Simulation study

In this section, we examine the performance of our estimators for IV in different types of noise settings. First, we study, through Monte Carlo simulations, the finite-sample properties of the estimator $\mathrm{PKRV}(Y)^{(n)(adj)}$ for IV in the presence of dependent and endogenous noise, and compare it with the kernel-based realized volatility estimator proposed by [14] using the modified Tukey-Hanning kernel with $p = 2$; this estimator is denoted as KB (kernel-based realized volatility). The KB is the only existing estimator for IV that is robust to special types of dependent and endogenous noise. Second, we study the performance of the estimator $\mathrm{TPKRV}(Z)^{(n)(adj)}$ for IV in the presence of dependent and endogenous noise and jumps. Here, we consider two noise-jump models: (i) A finite-dependent and endogenous microstructure noise process with a compound Poisson jump component, and (ii) a rounding error process [20] with compound Poisson jumps.

4.1. Simulation design

We consider the following two noise processes.
(1) In light of the facts of market microstructure noise empirically demonstrated in [8], we suppose an autoregressive and endogenous noise process ε_t given by the following dynamics:

$$\varepsilon_t = \varphi_1\, \varepsilon_{t-1} + \theta_1\, \Delta X_t + U_t, \tag{4.1}$$

where U_t is centered i.i.d. Gaussian with mean zero and variance $\mathrm{Var}(U_t) = w^2$, and $|\varphi_1| < 1$ and $|\theta_1| < 1$, where φ_1 and θ_1 are the regression coefficients of the noise and the coefficients of the endogenous noise, respectively. The microstructure noise series defined by Eq (4.1) is serially dependent and correlated with efficient returns. Furthermore, the efficient price X is assumed to follow a constant volatility Brownian process

$$dx_t = \mu_t\, dt + \sigma_t\, dw_t. \tag{4.2}$$

(2) In the presence of rounding error noise, following [20], the observed price can be expressed as

$$Y_i = \log\left(g \left\lfloor \frac{\exp(X_i + \varepsilon_i)}{g} \right\rfloor \right), \tag{4.3}$$

where

$$\varepsilon_i = \eta_i \log\left(\frac{g \left\lceil \frac{\exp(X_i)}{g} \right\rceil}{\exp(X_i)} \right)$$

and η_i is a Bernoulli variable (with the probabilities p_i and $1 - p_i$ taking values 1 and 0, respectively), with

$$p_i = \log\left(\frac{\exp(X_i)}{g\left\lfloor\frac{\exp(X_i)}{g}\right\rfloor}\right)\Big/\log\left(\frac{g\left\lceil\frac{\exp(X_i)}{g}\right\rceil}{g\left\lfloor\frac{\exp(X_i)}{g}\right\rfloor}\right),$$

where g is the rounding level.

We simulate data in the unit interval $[0, 1]$ and normalize 1 second to be $1/14700$, so that $[0, 1]$ is thought to span a typical trading day (4 hours and 5 minutes). The observed price process Y is generated by using an Euler scheme based on 14700 intervals. In the simulation, the parameters volatility σ_t and drift μ_t are always set to be $\sigma_t = 0.2/\sqrt{252}$ per day and $\mu_t = 0.03/252$ per day, respectively. The microstructure noise processes are generated from the models defined by Eqs (4.1) and (4.3), whose parameters are listed in Table 1, in which $g = 0.001, 0.01$ are mild rounding and serious rounding, respectively. Here, 1000 repetitions were run for each model.

Table 1. The model parameters.

Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
$\varphi_1 = 0$	$\varphi_1 = 0.1$	$\varphi_1 = 0.5$	$\varphi_1 = 0.9$	$\varphi_1 = 0.9$	$\varphi_1 = 0.9$		
$\theta_1 = 0$	$\theta_1 = -0.1$	$\theta_1 = -0.5$	$\theta_1 = -0.9$	$\theta_1 = -0.9$	$\theta_1 = -0.9$	g=0.001	g=0.01
$\omega = 0.0005$	$\omega = 0.0005$	$\omega = 0.0005$	$\omega = 0.0005$	$\omega = 0.0001$	$\omega = 0.001$		

(3) On the basis of the noise process setting above, we consider the following jump-diffusion process:

$$Z_i = X_i + \varepsilon_i + J_i, \tag{4.4}$$

where X_i follows Eq (4.2), and ε_i follows an autoregressive and endogenous noise model described in Eq (4.1) and the rounding error noise model described in Eq (4.3). For the compound Poisson jump process J_i, we set the intensity $\lambda = 3$ and the jump size $k_{\tau_i} \sim \mathrm{N}(0, \sigma_k^2)$. To clearly evaluate the impact of jumps on the estimator $\mathrm{TPKRV}(Z)^{(n)(adj)}$, we choose a relatively large jump size, $\sigma_k = 0.05$. In the presence of jumps, we proceed with the same steps as in the absence of jumps to determine the optimal window width k_n^*, except for redefining the variable d_t in the statistics $t_{(k_n, k_n')}$ as follows:

$$d_t = \mathrm{TPKRV}(Z)^{(n)(adj)}(k_n) - \mathrm{TPKRV}(Z)^{(n)(adj)}(k_n').$$

4.2. Results

4.2.1. Implementation of $\mathrm{PKRV}(Y)^{(n)(adj)}$

Table 2 reports the $t_{(k_n, k_n')}$ test statistics for the correlation time of models with various dependent and endogenous noise structures, as listed in Table 1. From these results, two main conclusions emerge. First, stronger serial dependence or endogeneity (or both) in the microstructure noise leads to a proportional lengthening of its correlation time. Second, the correlation time also increases with the variance of the noise. Both findings are intuitive and in line with theoretical expectations. Although this paper generally pursues a refined correlation time test, the subsequent analysis reveals that such refinement is only beneficial when it serves to reduce the probability of obtaining a negative IV estimate.

Table 2. Correlation time test for the estimator $\mathrm{PKRV}(Y)^{(n)(adj)}$.

k_n	1	2	3	4	5	6	7	8	9
$\varphi_1 = 0,\ \theta_1 = 0,\ \omega = 0.0005$	1.29	-0.39	1.31	0.96	-0.33	-0.68	-1.05	-0.32	0.66
$\varphi_1 = 0.1,\ \theta_1 = -0.1,\ \omega = 0.0005$	**257.99**	**25.23**	0.30	-0.25	-1.07	0.12	-0.94	-0.50	-1.05
k_n	7	8	9	10	20	30	40	50	60
$\varphi_1 = 0.5,\ \theta_1 = -0.5,\ \omega = 0.0005$	**9.83**	**4.78**	1.33	0.43	-0.73	-1.33	-1.84	-0.91	-0.18
k_n	10	15	25	35	40	45	50	55	65
$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.0005$	**294.03**	**163.82**	**48.89**	**14.60**	**9.21**	**5.91**	**3.76**	1.52	0.39
$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.0001$	**3.65**	**2.97**	**2.13**	**2.09**	**2.16**	1.95	1.73	1.54	1.38
$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.001$	**317.96**	**179.88**	**56.78**	**16.06**	**9.99**	**5.63**	**3.90**	**2.19**	0.83
g=0.001	1.49	**2.00**	**2.09**	1.45	1.15	0.84	0.51	0.21	-0.07
g=0.01	**3.00**	1.66	0.44	-0.03	-0.14	-0.27	-0.24	-0.41	-0.97

Notes: The $t_{(k_n, k_n')}$ statistics in bold denote rejection of the null hypothesis at the 5% significance level.

Figure 1 presents the volatility signature plots and the root mean square errors (RMSEs) of $\mathrm{PKRV}(Y)^{(n)(adj)}$ for various window widths k_n. Figure 1 shows that $\mathrm{PKRV}(Y)^{(n)(adj)}$ works quite well as an estimator of IV in the presence of dependent and endogenous market microstructure noise. The average estimated values of IV by $\mathrm{PKRV}(Y)^{(n)(adj)}$ are almost identical to the true value of the IV within a very large window width range, so the proposed estimator $\mathrm{PKRV}(Y)^{(n)(adj)}$ is not sensitive to the window width. We determine the optimal window width k_n^* by searching for the minimum of RMSE in the window width range that is larger than the correlation time. The optimal window widths k_n^* of the models are labeled in Figure 1.

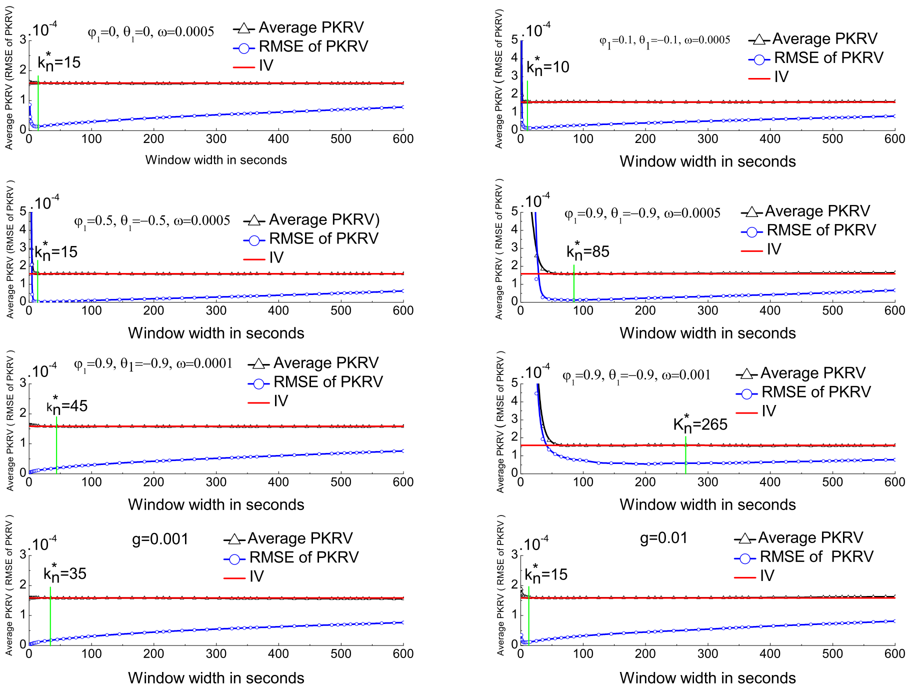

Figure 1. This figure shows the sample averages and the sample RMSEs of $\mathrm{PKRV}(Y)^{(n)(adj)}$, as a function of the window widths k_n. The optimal window widths k_n^* inferred from the minimum MSE criterion and the correlation time test are also labeled by the vertical line.

In this short section, we focus on assessing the central limit theory for $\mathrm{PKRV}(Y)^{(n)(adj)} - \mathrm{IV}$. The asymptotic variance of $\mathrm{PKRV}(Y)^{(n)(adj)}$ is written as Γ. This allows us to compute the asymptotic statistic

$$\frac{n^{1/6}(\mathrm{PKRV}(Y)^{(n)(adj)} - \mathrm{IV})}{\sqrt{\Gamma}} \xrightarrow{\mathcal{L}} \mathrm{N}(0,1). \tag{4.5}$$

In the following calculation of the statistic $PKRV(Y)^{(n)(adj)}$, we use the optimal window widths labelled in Figure 1. As seen in Figure 2, the standardized statistic defined in Eq (4.5) in various noise settings is essentially normally distributed even for a small sample size $n = 735$. This shows the importance of finite sample correction of the $\mathrm{PKRV}(Y)^{(n)(adj)}$ when dealing with small sample sizes.

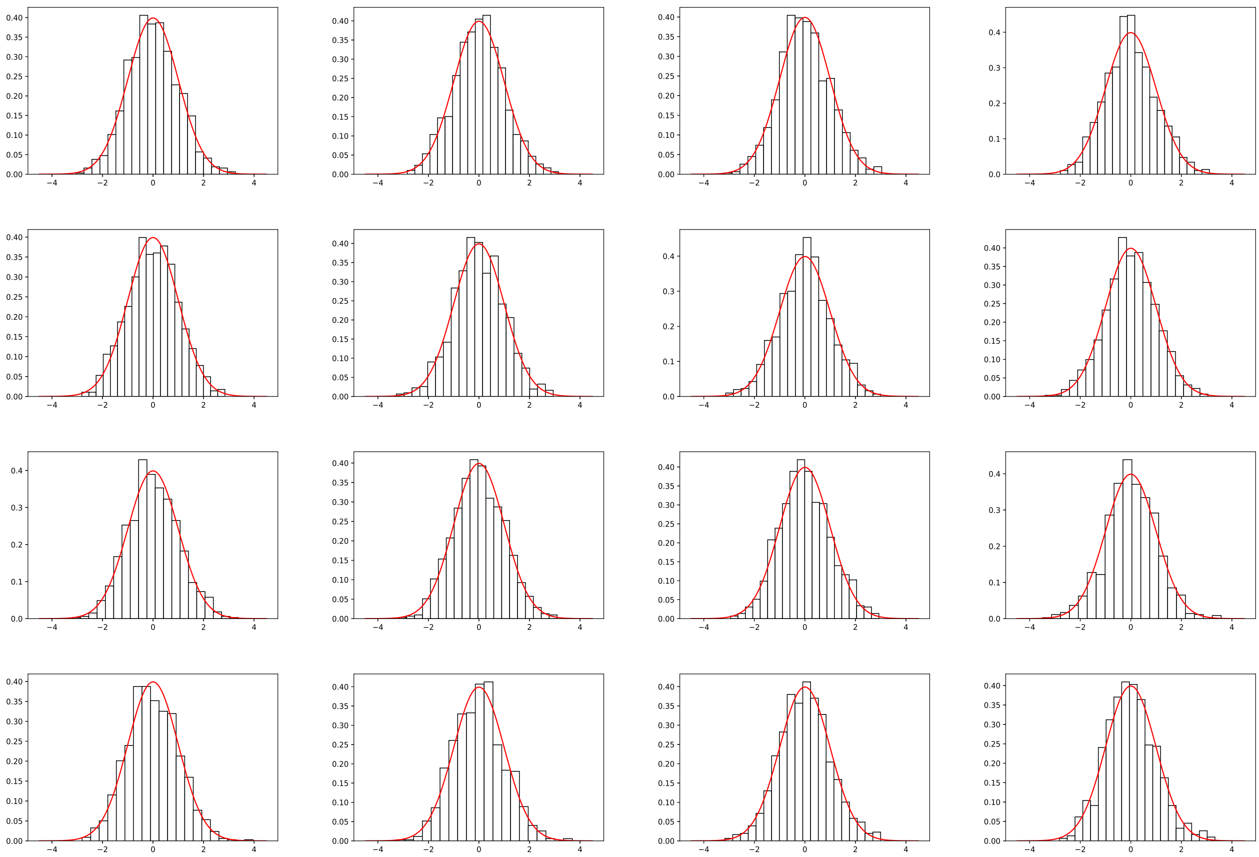

Figure 2. The various histograms of the statistics defined in Eq (4.5) are illustrated for microstructure noise Model 1 to Model 8 listed in Table 1 from top to bottom, respectively. In each line, the first data was computed with $n = 14700$, whereas for the second one, we used $n = 735$. For comparison, the red lines show the graph of the standard normal density.

Table 3 shows the Monte Carlo results for mean and standard deviation of both $\mathrm{PKRV}(Y)^{(n)(adj)} - \mathrm{IV}$ and $\mathrm{KB}(Y)^{(n)} - \mathrm{IV}$. It shows that $\mathrm{PKRV}(Y)^{(n)(adj)}$ works quite well as an estimator of IV in the presence of dependent and endogenous noise, and in the rounding error noise setting, since both the bias and standard deviation are rather small, at least for sample sizes larger than $n = 735$. For the i.i.d. noise setting, both $\mathrm{PKRV}(Y)^{(n)(adj)}$ and $\mathrm{KB}(Y)^{(n)}$ can exhibit the same excellent performance. $\mathrm{PKRV}(Y)^{(n)(adj)}$ provides much better finite sample properties than $\mathrm{KB}(Y)^{(n)}$ with the increase in the dependence and endogeneity of noise, and also with more serious rounding. Taken overall, the finite sample bias of $\mathrm{PKRV}(Y)^{(n)(adj)}$ is much smaller than that of $\mathrm{KB}(Y)^{(n)}$. Table 3 shows that the estimator $\mathrm{KB}(Y)^{(n)}$ of IV is invalid for the strongly dependent and strongly endogenous microstructure noise case. However, $\mathrm{PKRV}(Y)^{(n)(adj)}$ works quite well even for such strongly dependent and strongly endogenous microstructure noise, for example, $\varphi_1 = 0.9$, $\theta_1 = -0.9$, and $\omega = 0.001$.

Table 3. The Monte Carlo results for the mean and standard deviation of both $\mathrm{PKRV}(Y)^{(n)(adj)} - \mathrm{IV}$ and $\mathrm{KB}(Y)^{(n)} - \mathrm{IV}$. In the implementation of $\mathrm{PKRV}(Y)^{(n)}$, we used the optimal window widths k_n^*.

	$\times 10^{-4}$							
	$\varphi_1 = 0,\ \theta_1 = 0,\ \omega = 0.0005$		$\varphi_1 = 0.1,\ \theta_1 = -0.1,\ \omega = 0.0005$		$\varphi_1 = 0.5,\ \theta_1 = -0.5,\ \omega = 0.0005$		$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.0005$	
n	Mean	Std	Mean	Std	Mean	Std	Mean	Std
$\mathrm{PKRV}(Y)^{(n)(adj)}$								
735	0.004	0.3011	0.007	0.2927	0.020	0.3539	-0.011	0.4158
1470	-0.004	0.2244	-0.001	0.3376	0.004	0.2732	-0.007	0.3750
2940	-0.003	0.1982	-0.002	0.2440	0.001	0.2368	-0.007	0.3587
4900	-0.007	0.1714	-0.009	0.1912	0.002	0.2033	-0.004	0.3598
7350	0.000	0.1558	0.007	0.1739	-0.004	0.1879	-0.005	0.3569
14700	-0.003	0.1347	0.008	0.1433	-0.001	0.1777	-0.005	0.3550
$\mathrm{KB}(Y)^{(n)}$								
735	0.011	0.2150	0.015	0.2182	0.008	0.2308	0.063	0.3542
1470	0.001	0.1644	0.013	0.1939	0.008	0.2055	0.099	0.3384
2940	0.001	0.1360	0.012	0.1803	0.012	0.1897	0.139	0.3320
4900	0.001	0.1202	0.012	0.1739	0.022	0.1834	0.164	0.3289
7350	0.004	0.1107	0.013	0.1708	0.033	0.1803	0.177	0.3257
14700	0.003	0.1012	0.014	0.1676	0.052	0.1803	0.193	0.3257
	$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.0001$		$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.001$		$g = 0.001$		$g = 0.01$	
n	Mean	Std	Mean	Std	Mean	Std	Mean	Std
$\mathrm{PKRV}(Y)^{(n)(adj)}$								
735	-0.016	0.2274	-0.007	0.6161	-0.005	0.1930	0.003	0.1622
1470	-0.009	0.2052	0.002	0.6066	-0.004	0.1889	0.003	0.1490
2940	-0.005	0.1932	0.003	0.5973	-0.000	0.1770	0.011	0.1291
4900	-0.005	0.1929	0.008	0.5891	-0.001	0.1789	0.011	0.1262
7350	-0.006	0.1953	0.007	0.5897	-0.002	0.1789	0.011	0.1261
14700	-0.005	0.1931	0.006	0.5895	-0.000	0.1766	0.012	0.1222
$\mathrm{KB}(Y)^{(n)}$								
735	-0.008	0.2119	0.204	0.4174	0.003	0.1518	0.004	0.1581
1470	-0.007	0.1992	0.368	0.3953	0.006	0.1265	0.018	0.1233
2940	-0.006	0.1897	0.534	0.3890	0.005	0.1107	0.030	0.1012
4900	-0.006	0.1866	0.632	0.3858	0.004	0.1044	0.036	0.0949
7350	-0.005	0.1834	0.689	0.3826	0.004	0.1012	0.044	0.0885
14700	-0.005	0.1834	0.754	0.3858	0.004	0.0980	0.053	0.0822

A drawback of the $\mathrm{PKRV}(Y)^n$ is that it may produce a negative estimate of IV. When this occurs, a common remedy is to truncate the estimate at zero (the same technique is used in [14]). Table 4 shows that such negative estimates are extremely rare in our simulations. More importantly, for the selected optimal window width k_n^*, no negative IV estimate was observed across the 8000 samples. Since $\mathrm{PKRV}(Y)^n$ is unbiased and consistent for IV under the optimal window width k_n^*, the probability of obtaining a negative estimate vanishes asymptotically.

Table 4. Number of negative IVs for the estimator $\mathrm{PKRV}(Y)^{(n)(adj)}$.

	[1, 15]	[25, 125]	[145, 305]	[325, 465]	[485, 600]
$\varphi_1 = 0,\ \theta_1 = 0,\ \omega = 0.0005$	38/10000 (0/1000)	0/10000	0/10000	0/10000	0/10000
$\varphi_1 = 0.1,\ \theta_1 = -0.1,\ \omega = 0.0005$	0/10000 (0/1000)	0/10000	0/10000	0/10000	0/10000
$\varphi_1 = 0.5,\ \theta_1 = -0.5,\ \omega = 0.0005$	0/10000 (0/1000)	0/10000	0/10000	0/10000	2/10000
$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.0005$	0/10000	0/10000 (0/1000)	0/10000	0/10000	0/10000
$\varphi_1 = 0.9,\ \theta_1 = -0.9, \omega = 0.0001$	0/10000	0/10000 (0/1000)	0/10000	0/10000	0/10000
$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.001$	0/10000	348/10000	14/10000 (0/1000)	12/10000	45/10000
$g = 0.001$	0/10000	0/10000 (0/1000)	0/10000	0/10000	7/10000
$g = 0.01$	0/10000 (0/1000)	0/10000	0/10000	0/10000	0/10000

Notes: The numbers in the squared brackets denote the k_n-value range, in which we randomly choose 10 values of k_n. The numerators denote the number of days that the $\mathrm{PKRV}(Y)^{(n)(adj)}$ produced a negative estimate. The denominators denote the total number of days. The numbers in parentheses give the results where $\mathrm{PKRV}(Y)^{(n)(adj)}$ produced non-negative estimates for the optimal window width k_n^*.

4.2.2. Implementation of $\mathrm{TPKRV}(Z)^{(n)(adj)}$

Before assessing the performance of the estimator $\mathrm{TPKRV}(Z)^{(n)(adj)}$ for IV, we first need to choose a suitable threshold function $\Theta(\Delta_n) = C\Delta_n^\alpha$ for any $\alpha \in (0, 1/3)$ and $C \in R$. The choice of C is more critical than that of α, as it depends on factors such as the noise level, jump size, and volatility of the efficient price process. We set $\alpha = 0.25$ for $\mathrm{TPKRV}(Z)^{(n)(adj)}$. Figure 3 plots the relative bias of the estimator $\mathrm{TPKRV}(Z)^{(n)(adj)}$ for IV as a function of the threshold parameter C. As shown, the estimator is robust to C over the interval $[0.5, 1.0]$, with the relative bias remaining rather small across various market microstructure noise settings and in the presence of large jumps. Accordingly, we adopt $C = 0.7$ for both our simulation and empirical studies.

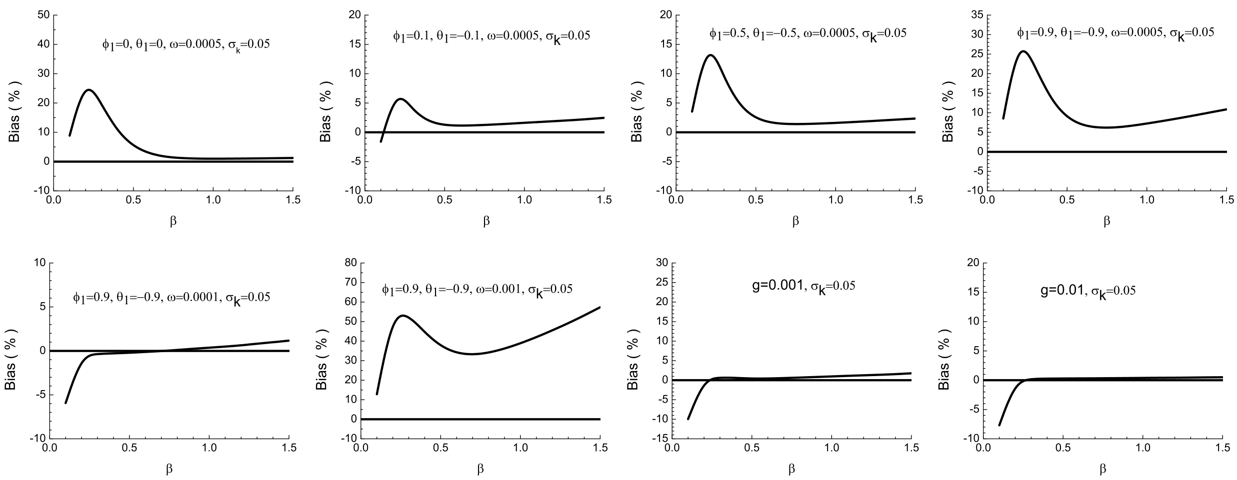

Figure 3. Relative biases of the estimator $\mathrm{TPKRV}(Z)^{(n)(adj)}$ for IV as a function of the threshold parameter c.

Table 5 presents the $t_{(k_n,k'_n)}$ test statistics for models using $\mathrm{TPKRV}(Z)^{(n)(adj)}$. For comparison, models with jumps and their jump-free counterparts were generated using the same continuous price process in our simulations. Table 5 shows that the correlation time changes little after the introduction of jumps, relative to Table 2. This is reasonable, because they share the same dependent noise structure.

Table 5. The correlation time test for the estimator $\mathrm{TPKRV}(Z)^{(n)(adj)}$.

k_n	13	14	15	16	17	18	19	20	25
$\varphi_1=0,\ \theta_1=0,\ \omega=0.0005, \sigma_k=0.05$	**6.89**	**4.82**	**3.05**	**2.54**	1.78	0.72	-0.23	0.64	0.47
$\varphi_1=0.1,\ \theta_1=-0.1,\ \omega=0.0005, \sigma_k=0.05$	**6.31**	**4.53**	**2.13**	0.99	0.25	0.60	0.50	-0.09	-1.01
$\varphi_1=0.5,\ \theta_1=-0.5,\ \omega=0.0005, \sigma_k=0.05$	**19.55**	**13.31**	**9.50**	**6.62**	**5.14**	**4.27**	**3.43**	**2.71**	0.01
$\varphi_1=0.9,\ \theta_1=-0.9,\ \omega=0.0001, \sigma_k=0.05$	**2.71**	**2.47**	**2.20**	1.93	1.67	1.38	1.18	1.06	0.26
k_n	45	50	55	60	65	66	67	68	69
$\varphi_1=0.9,\ \theta_1=-0.9,\ \omega=0.0005,\ \sigma_k=0.05$	**18.06**	**11.79**	**6.65**	**3.54**	**1.99**	1.86	1.70	1.56	1.48
k_n	185	190	195	200	205	210	215	220	225
$\varphi_1=0.9,\ \theta_1=-0.9,\ \omega=0.001,\ \sigma_k=0.05$	**5.26**	**3.70**	**3.77**	**2.60**	**3.07**	**2.55**	**2.26**	1.35	**2.00**
k_n	11	12	13	14	15	16	17	18	19
$g=0.001,\ \sigma_k=0.05$	**3.85**	**3.30**	**3.62**	**2.17**	**2.27**	1.45	-0.23	0.60	-0.44
k_n	1	2	3	4	5	6	7	8	9
$g=0.01,\ \sigma_k=0.05$	**13.33**	**8.20**	**4.73**	**2.72**	**2.26**	1.70	1.42	1.69	1.60

Notes: The $t_{(k_n,k'_n)}$ statistics in bold denote rejection of the null hypothesis at the 5% significance level.

Figure 4 presents the volatility signature plots and the RMSEs of $\mathrm{TPKRV}(Z)^{(n)(adj)}$ for a range of window widths k_n. According to Table 5 and Figure 4, on the basis of the selection criterion for the optimal window width k_n^* described in Section 3.3, we obtain the optimal window widths k_n^*, as labelled in Figure 4. The average IV estimate obtained by $\mathrm{TPKRV}(Z)^{(n)(adj)}$ is close to the true IV when $k_n = k_n^*$. Figure 4 shows that $\mathrm{TPKRV}(Z)^{(n)(adj)}$ performs quite well as an estimator of IV in settings with dependent and endogenous noise and jumps, provided that k_n is near the optimal window width. The proposed estimator $\mathrm{TPKRV}(Z)^{(n)(adj)}$ is still not sensitive to the window width.

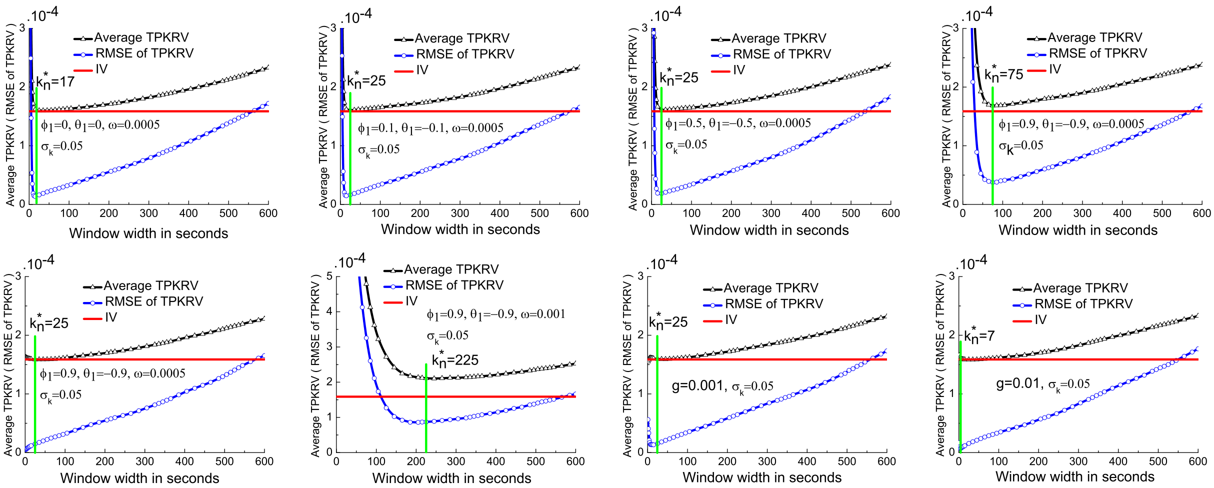

Figure 4. The figure shows the sample averages and the sample RMSEs of $\mathrm{TPKRV}(Z)^{(n)(adj)}$ as a function of the window widths k_n. The sampling intervals are always chosen to be $\Delta_n = 1$ second for the all price series. The optimal window widths k_n^* inferred from the minimum MSE criterion and the correlation time test are also labeled by the vertical line.

Table 6 reports the Monte Carlo results for the mean and standard deviation of $\mathrm{TPKRV}(Z)^{(n)(adj)} - \mathrm{IV}$. The results indicate that $\mathrm{TPKRV}(Z)^{(n)(adj)}$ works quite well as an estimator of IV in the presence of dependent and endogenous noise and jumps, with both the bias and standard deviation remaining rather small for sample sizes as small as $n = 735$. Only in the case with a large jump and strongly dependent and endogenous noise do the bias and standard deviation become somewhat larger. Even in this worst-case scenario, represented by the model with ($\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.001, \sigma_k = 0.05$), approximately 99.3% ($\frac{75-0.524}{75} \approx 99.3\%$) of the jump variance has been filtered. Therefore, we conclude that $\mathrm{TPKRV}(Z)^{(n)(adj)}$ is an efficient estimator for IV in the presence of dependent and endogenous noise and jumps.

Table 6. The Monte Carlo results for the mean and standard deviation of $\mathrm{TPKRV}(Z)^{(n)(adj)} - \mathrm{IV}$. In the implementation of $\mathrm{TPKRV}(Z)^{(n)(adj)}$, we used the optimal window widths k_n^*.

$\times 10^{-4}$								
	$\varphi_1 = 0,\ \theta_1 = 0$ $\omega = 0.0005,\ \sigma_k = 0.05$		$\varphi_1 = 0.1,\ \theta_1 = -0.1$ $\omega = 0.0005,\ \sigma_k = 0.05$		$\varphi_1 = 0.5,\ \theta_1 = -0.5$ $\omega = 0.0005,\ \sigma_k = 0.05$		$\varphi_1 = 0.9,\ \theta_1 = -0.9$ $\omega = 0.0005,\ \sigma_k = 0.05$	
n	Mean	Std	Mean	Std	Mean	Std	Mean	Std
735	0.239	0.2254	-0.009	0.2142	0.034	0.2204	0.392	0.2690
1470	0.206	0.1986	0.052	0.1960	0.134	0.2077	0.394	0.2857
2940	0.111	0.1794	0.051	0.1862	0.135	0.2010	0.326	0.3150
4900	0.079	0.1684	0.028	0.1777	0.051	0.1912	0.208	0.3352
7350	0.042	0.1523	0.018	0.1713	0.032	0.1872	0.135	0.3475
14700	0.019	0.1430	0.0190	0.1639	0.022	0.1850	0.098	0.3632
	$\varphi_1 = 0.9,\ \theta_1 = -0.9$ $\omega = 0.0001,\ \sigma_k = 0.05$		$\varphi_1 = 0.9,\ \theta_1 = -0.9$ $\omega = 0.001,\ \sigma_k = 0.05$		$g = 0.001$ $\sigma_k = 0.05$		$g = 0.01$ $\sigma_k = 0.05$	
n	Mean	Std	Mean	Std	Mean	Std	Mean	Std
735	-0.030	0.1806	0.645	0.3087	-0.036	0.1943	-0.032	0.1477
1470	-0.011	0.1605	0.856	0.3604	-0.002	0.1755	-0.003	0.1062
2940	-0.006	0.1481	0.825	0.4569	0.012	0.1630	0.000	0.10112
4900	-0.007	0.1525	0.698	0.5374	0.007	0.1619	0.003	0.0958
7350	-0.004	0.1503	0.590	0.5924	0.006	0.1577	0.003	0.08968
14700	-0.000	0.1484	0.524	0.6952	0.009	0.1540	0.005	0.0848

Table 7 presents the number of negative IV estimates obtained by $\mathrm{TPKRV}(Z)^{(n)}$. Overall, the occurrence of negative IV estimates is very rare. We also find no negative IV estimates for the optimal window width k_n^* across the 8000 samples.

Table 7. Number of negative IVs for the estimator $\text{TPKRV}(Z)^{(n)(adj)}$.

	[1, 15]	[25, 125]	[145, 305]	[325, 465]	[485, 600]
$\varphi_1 = 0,\ \theta_1 = 0,\ \omega = 0.0005,\ \sigma_k = 0.05$	0/10000(0/1000)	0/10000	0/10000	0/10000	6/10000
$\varphi_1 = 0.1,\ \theta_1 = -0.1,\ \omega = 0.0005,\ \sigma_k = 0.05$	0/10000	0/10000(0/1000)	0/10000	0/10000	0/10000
$\varphi_1 = 0.5,\ \theta_1 = -0.5,\ \omega = 0.0005,\ \sigma_k = 0.05$	0/10000	0/10000(0/1000)	0/10000	0/10000	0/10000
$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.0005,\ \sigma_k = 0.05$	0/10000	0/10000(0/1000)	0/10000	0/10000	0/10000
$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.0001,\ \sigma_k = 0.05$	0/10000	0/10000(0/1000)	0/10000	0/10000	4/10000
$\varphi_1 = 0.9,\ \theta_1 = -0.9,\ \omega = 0.001,\ \sigma_k = 0.05$	0/10000	0/10000	0/10000 (0/1000)	0/10000	4/10000
$g = 0.001,\ \sigma_k = 0.05$	0/10000	0/10000(0/1000)	0/10000	0/10000	0/10000
$g = 0.01,\ \sigma_k = 0.05$	0/10000(0/1000)	0/10000	0/10000	0/10000	0/10000

Notes: The numbers in the squared brackets denote the k_n-value range, in which we randomly choose 10 values of k_n. The numerators denote the number of days that the $\text{TPKRV}(Z)^{(n)(adj)}$ produced a negative estimate. The denominators denote the total number of days. The numbers in parentheses give the results where the $\text{TPKRV}(Z)^{(n)(adj)}$ produced non-negative estimates for the optimal window widths k_n^*.

5. Empirical study

We put our theoretical results to work in the empirical analysis of four constituent stocks from the Shanghai Stock Exchange (SSE) 50 Index. The sample period spans from April 2, 2020 to April 29, 2022, covering 485 trading days. The data are transaction prices from the China Stock Market & Accounting Research (CSMAR) Level 2 database. We discarded transactions outside the periods 9:30–11:30 a.m. and 1:00–3:00 p.m. to focus on the continuous trading session of the Chinese stock market. We also removed entries with nonpositive prices. The opening and closing auction periods are excluded because their price formation mechanisms differ from those of continuous trading and may introduce additional microstructure effects that are not the focus of this paper. The four stocks, namly China National Petroleum Corporation (CNPC), Industrial and Commercial Bank of China (ICBC), China Union Communication (CUCC), and Kweichow Moutai (KCMT), are highly actively traded; their average number of transactions per second is about 2 to 4.

Equidistant price data were constructed from raw (irregularly spaced) price data using the previous tick method. Summary statistics for the intraday returns are reported in Table 8, and their autocorrelation and partial autocorrelation functions are presented in Figure 5. Table 8 shows that the high-frequency intraday return series are stationary and dependent. As seen in Figure 5, the autocorrelation and partial autocorrelation coefficients decay toward zero. Therefore, the observed high-frequency intraday return series are complex and should be approximately modeled by an autoregressive moving average (ARMA) model rather than a first-order moving average (MA(1)) model. This illustrates that the assumption of dependent and endogenous market microstructure noise is necessary. On the basis of the excellent performance of $\text{PKRV}(Y)^{(n)(adj)}$ and $\text{TPKRV}(Z)^{(n)(adj)}$, and the validity of the $t_{(k_n, k_n')}$ test exhibited in the simulation studies, we apply the proposed estimators $\text{PKRV}(Y)^{(n)(adj)}$ and $\text{TPKRV}(Z)^{(n)(adj)}$ and the $t_{(k_n, k_n')}$ test to the empirical analysis of the four stocks from the SSE 50 Index.

Table 8. Summary information for high-frequency intraday returns.

	CNPC	ICBC	CUCC	KCMT
Mean	0.0000025	-0.0000011	-0.0000053	0.0000085
Std	0.106	0.105	0.105	0.036
Skewness	0.0309	-0.0260	-0.0052	-0.3538
Kurtosis	5.3614	2.0721	2.4976	8269.5
ADF	-201.2***	-195.8***	-209.2***	-205.7***
$Q_{LB}(5)$	801513***	814954***	698844***	579250***
$Q_{LB}(10)$	802143***	895950***	700157***	602244***
$Q_{LB}(20)$	804072***	900558***	703182***	612066***

Notes: This table presents the summary information for log-returns of CNPC, ICBC, CUCC, and KCMT. The t-statistics marked with *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

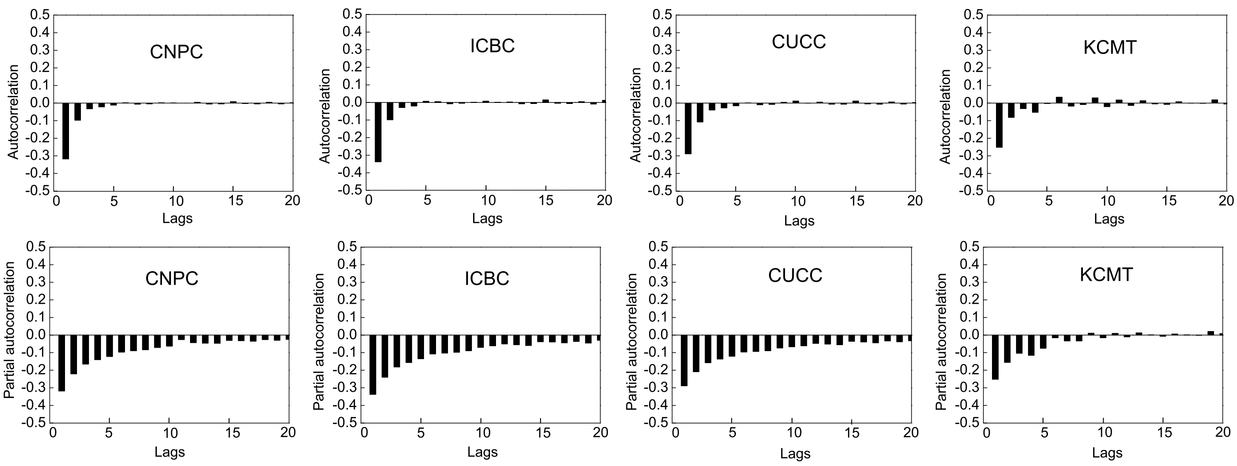

Figure 5. The figure illustrates the log-return autocorrelograms and the log-return partial autocorrelograms from transactions based on the sampling interval $\Delta_n = 1$ for CNPC, ICBC, CUCC, KCMT. The sample period is from April 2, 2020 to April 29, 2022, consisting of 485 trading days.

Table 9 presents the t-values of the correlation time test for the four stocks. For CNPC, $k_n = 30$ passes the $t_{(k_n,k_n')}$ test but, subsequently, $k_n = 130$ does not pass the $t_{(k_n,k_n')}$ test. As shown in Table 8, the $Q_{LB}(5)$, $Q_{LB}(10)$, and $Q_{LB}(20)$ statistics indicate that the intraday return series of CNPC are correlated, at least for lags of up to 20. Generally, choosing a larger k_n as the correlation time yields more reliable results. Therefore, we do not choose the correlation time of 30. We set the step size to $\Delta s = 10$ and repeat the $t_{(k_n,k_n')}$ test, and then choose 140 as the correlation time. Figure 6 presents the volatility signature plots and standard deviations of $\text{TPKRV}(Z)^{(n)(adj)}$ for CNPC, ICBC, CUCC, and KCMT as a function of various window widths k_n. Table 10 presents the number of negative IV estimates from $\text{TPKRV}(Z)^{(n)(adj)}$ for CNPC, ICBC, CUCC, and KCMT. Considering the correlation time, the minimum standard deviation of $\text{TPKRV}(Z)^{(n)(adj)}$, as shown in Figure 6, and the number of negative IV estimates presented in Table 10, we choose the optimal window width $k_n^* = 145$ for CNPC. Following the same steps, we choose the optimal window widths for ICBC, CUCC, and KCMT, and these optimal widths are labeled in Figure 6.

Table 9. Correlation time test of CNPC, ICBC, CUCC, and KCMT with TPKRV$(Z)^{(n)(adj)}$.

k_n	10	30	50	70	90	110	130	150	170
CNPC	**3.52**	-1.43	-1.10	-0.93	0.30	1.08	2.17	0.52	-0.03
ICBC	**18.94**	-1.01	0.03	**2.79**	0.13	**2.94**	1.45	-0.84	-1.58
CUCC	**3.31**	**2.33**	**2.86**	**4.90**	**3.83**	**5.36**	**4.57**	**3.51**	1.60
KCMT	**-9.97**	**3.32**	**4.63**	**2.64**	0.70	0.12	0.11	1.32	**2.70**

Notes: The $t_{(k_n,180)}$ statistics in bold denote rejection of the null hypothesis at the 5% significance level.

Table 10. Number of negative IVs for the TPKRV$^{(n)}$ estimator for CNPC, ICBC, CUCC and KCMT.

	[1, 15]	[25, 125]	[145, 305]	[325, 465]	[485,600]
CNPC(145)	62/4850	48/4850	5/4850(0/485)	19/4850	38/4850
ICBC(180)	181/4850	158/4850	34/4850(0/485)	35/4850	65/4850
CUCC(185)	71/4850	18/4850	6/4850(0/485)	21/4850	53/4850
KCMT(80)	0/4850	0/4850(0/485)	0/4850	0/4850	10/4850

Notes: The numbers in the squared brackets denote the k_n-value range, in which we randomly choose 10 values of k_n. The numerators denote the number of days that the TPKRV$(Z)^{(n)(adj)}$ produced a negative estimate. The denominators denote the total number of days. The numbers in parentheses give the results where the TPKRV$(Z)^{(n)(adj)}$ produced non-negative estimates for the optimal window widths k_n^*.

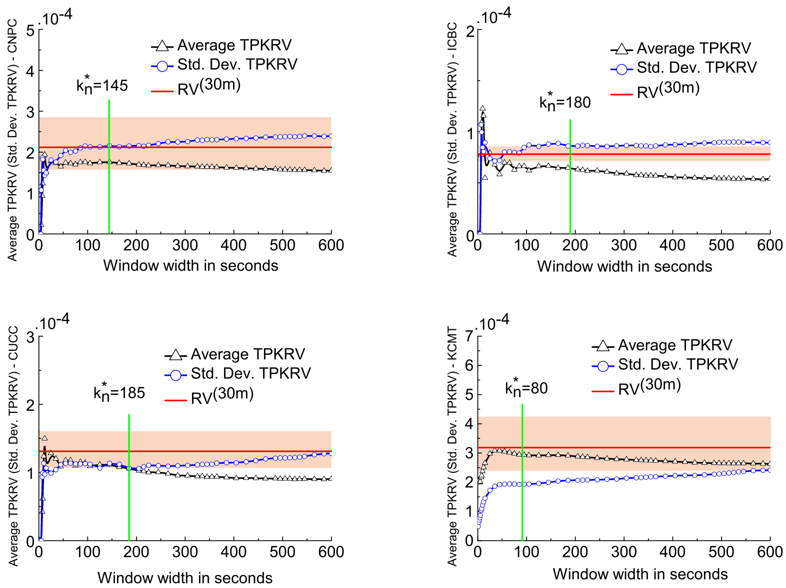

Figure 6. The figure shows the sample averages and the sample standard deviations of TPKRV$(Z)^{(n)(adj)}$ for CNPC, ICBC, CUCC, and KCMT, as a function of the window widths k_n. The signature plots of the sample standard deviation of TPKRV$^{(n)(adj)}$ can be viewed as an approximation of the MSE signature plots, because the TPKRV$^{(n)(adj)}$ is a conditionally unbiased estimator of IV. The sampling intervals are always chosen to be $\Delta_n = 1$. The shaded area about the average $RV^{(30m)}$ represents an approximate 95% confidence interval for the average IV. The optimal window widths k_n^* inferred from the minimum MSE criterion and correlation time test are also labeled by the vertical line.

As shown in Table 10, there are no negative IV estimates in the nearly 2000 samples using the optimal window widths. Therefore, precisely determining the optimal window width k_n^* is very helpful in reducing the probability of negative IV estimates for the estimators $\mathrm{TPKRV}(Z)^{(n)(adj)}$ and $\mathrm{PKRV}(Y)^{(n)(adj)}$.

The probability of a negative IV estimate has rarely been systematically studied in the pre-averaging or kernel literature. In our simulations (8000 replications) and empirical study (2000 daily observations), no negative IV estimate occurred when the optimal window width was used, an observation that supports the practical relevance of our window selection procedure.

As shown in Figure 6, the estimated IV of $\mathrm{TPKRV}(Z)^{(n)(adj)}$ at the optimal window width is very close to $\bar{R}V^{(30m)}$, which is an appropriate target for average volatility and should be approximately unbiased for the average IV; moreover, the estimator $\mathrm{TPKRV}(Z)^{(n)(adj)}$ is insensitive to the window width. Table 11 also shows that $\mathrm{PKRV}(Y)^{(n)(adj)}$ and $\mathrm{TPKRV}(Z)^{(n)(adj)}$ have excellent finite sample properties, at least for sample sizes larger than 735 and 7350, respectively.

Table 11. The mean and standard deviation of $\mathrm{TPKRV}^{(n)}$ for CNPC, ICBC, CUCC, and KCMT, using the optimal window widths k_n^* in the implementation.

	$\times 10^{-4}$							
	CNPC ($k_n^* = 145$)		ICBC ($k_n^* = 180$)		CUCC ($k_n^* = 185$)		KCMT ($k_n^* = 80$)	
n	Mean	Std	Mean	Std	Mean	Std	Mean	Std
$\mathrm{PKRV}(Y)^{(n)(adj)}$								
735	2.066	3.2278	0.705	1.1238	1.158	1.6719	3.928	12.7706
1470	2.071	3.2323	0.706	1.1189	1.164	1.6786	3.905	12.2363
2940	2.088	3.2216	0.705	1.1178	1.179	1.6811	3.899	12.1017
4900	2.089	3.2286	0.701	1.1029	1.179	1.6753	3.894	12.0771
7350	2.090	3.2303	0.700	1.1043	1.175	1.6729	3.900	12.1277
14700	2.092	3.2289	0.701	1.1056	1.175	1.6764	3.898	12.1009
$\mathrm{TPKRV}(Z)^{(n)(adj)}$								
735	1.119	0.8599	0.556	0.5183	0.833	0.5962	1.930	0.8138
1470	1.248	1.1417	0.601	0.6728	0.899	0.7559	2.235	1.0895
2940	1.409	1.3761	0.626	0.7042	0.955	0.8190	2.502	1.3467
4900	1.542	1.6502	0.643	0.7925	0.997	0.8881	2.675	1.5384
7350	1.624	1.8080	0.648	0.8048	1.017	0.9561	2.796	1.6868
14700	1.758	2.1577	0.664	0.8658	1.059	1.0612	2.964	1.9328

Our estimators $\mathrm{TPKRV}(Z)^{(n)(adj)}$ and $\mathrm{PKRV}(Y)^{(n)(adj)}$ perform well in the empirical analyses. They not only yield accurate IV estimates but are also insensitive to the window width and sampling frequency. Overall, the $\mathrm{TPKRV}(Z)^{(n)(adj)}$ is valid in the empirical analyses and robust to various noises and jumps.

6. Conclusions

In high-frequency financial data, the estimators for IV were usually established based on the i.i.d. noise assumption, which deviates significantly from the financial market. In this paper, we develop the estimator PKRV$(Y)^{(n)}$ for IV, which allows the market microstructure noise to be serially dependent and endogenous, while only assuming that the noise process is difference-stationary and has finite correlations. The correlation time test statistics $t_{(k_n, k'_n)}$ and the minimum MSE criterion are shown to be valid for determining the optimal window width in the simulation experiments. The estimator TPKRV$(Z)^{(n)}$ for IV is developed by combining PKRV$(Y)^{(n)}$ with the threshold approach for the price process with jumps. The consistency and asymptotic normality of these two estimators, PKRV$(Y)^{(n)}$ and TPKRV$(Z)^{(n)}$, have been well established. The simulation and empirical studies show that PKRV$(Y)^{(n)}$ and TPKRV$(Z)^{(n)}$ provide excellent finite-sample performance in comparison with the alternative estimator. They produce purely positive IV estimates at the optimal window width and are insensitive to sampling frequency and to the choice of k_n^*. This also indirectly suggests that our procedure for determining the optimal window width is feasible.

Use of generative AI tools declaration

The author declares he has not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the National Natural Science Foundation, No. 72201090 and Faculty Start-up Fund of Hubei University of Technology, No. XJ2022007601.

Conflict of interest

The author declares no conflict of interest.

References

1. T. G. Andersen, T. Bollerslev, Answering the skeptics: yes, standard volatility models do provide accurate forecasts, *Int. Econ. Rev.*, **39** (1998), 885–905. https://doi.org/10.2307/2527343
2. O. E. Barndorff-Nielsen, N. Shephard, Estimating quadratic variation using realized variance, *J. Appl. Econom.*, **17** (2002), 457–477. https://doi.org/10.1002/jae.691
3. R. M. Leushuis, N. Petkov, Advances in forecasting realized volatility: a review of methodologies, *Financ. Innov.*, **12** (2026), 14. https://doi.org/10.1186/s40854-025-00809-5
4. K. Yang, G. Zhang, D. Wang, On smooth transition interval autoregressive models, *J. Bus. Econ. Stat.*, **44** (2026), 613–625. https://doi.org/10.1080/07350015.2025.2546455
5. H. Li, Q. Zhang, K. Yang, Bayesian forecasting for a logistic mixture double autoregressive model, *J. Forecast.*, **45** (2026), 1665–1680. https://doi.org/10.1002/for.70109
6. H. A. Ahmed, Dynamics among the term spread, stock market volatility forecast, financial market risk and oil price: an empirical analysis, *Financ. Innov.*, **12** (2026), 24. https://doi.org/10.1186/s40854-025-00862-0
7. B. Korkusuz, M. Sahiner, Coin impact on cross-crypto realized volatility and dynamic cryptocurrency volatility connectedness, *Financ. Innov.*, **11** (2025), 129. https://doi.org/10.1186/s40854-025-00881-x

8. P. R. Hansen, A. Lunde, Realized variance and market microstructure noise, *J. Bus. Econ. Stat.*, **24** (2006), 127–161. https://doi.org/10.1198/073500106000000071

9. F. X. Diebold, G. Strasser, On the correlation structure of microstructure noise: a financial economic approach, *Rev. Econ. Stud.*, **80** (2013), 1304–1337. https://doi.org/10.1093/restud/rdt008

10. T. G. Andersen, I. Archakov, G. Cebiroglu, N. Hautsch, Local mispricing and microstructural noise: a parametric perspective, *J. Econom.*, **230** (2022), 510–534. https://doi.org/10.1016/j.jeconom.2021.06.006

11. L. Zhang, P. A. Mykland, Y. Aït-Sahalia, A tale of two time scales: determining integrated volatility with noisy high-frequency data, *J. Amer. Statist. Assoc.*, **100** (2005), 1394–1411. https://doi.org/10.1198/016214505000000169

12. L. Zhang, Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach, *Bernoulli*, **12** (2006), 1019–1043. https://doi.org/10.3150/bj/1165269149

13. B. Zhou, High-frequency data and volatility in foreign-exchange rates, *J. Bus. Econ. Stat.*, **14** (1996), 45–52.

14. O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, N. Shephard, Designing realized kernels to measure the ex post variation of equity prices in the presence of noise, *Econometrica*, **76** (2008), 1481–1536. https://doi.org/10.3982/ECTA6495

15. Y. Aït-Sahalia, P. A. Mykland, L. Zhang, How often to sample a continuous-time process in the presence of market microstructure noise, *Rev. Financ. Stud.*, **18** (2005), 351–416. https://doi.org/10.1093/rfs/hhi016

16. D. C. Xiu, Quasi-maximum likelihood estimation of volatility with high frequency data, *J. Econom.*, **159** (2010), 235–250. https://doi.org/10.1016/j.jeconom.2010.07.002

17. R. Da, D. C. Xiu, When moving-average models meet high-frequency data: uniform inference on volatility, *Econometrica*, **89** (2021), 2787–2825. https://doi.org/10.3982/ECTA15593

18. S. Clinet, Y. Potiron, Testing if the market microstructure noise is fully explained by the informational content of some variables from the limit order book, *J. Econom.*, **209** (2019), 289–337. https://doi.org/10.1016/j.jeconom.2019.01.004

19. M. Podolskij, M. Vetter, Bipower-type estimation in a noisy diffusion setting, *Stochastic Process. Appl.*, **119** (2009), 2803–2831. https://doi.org/10.1016/j.spa.2009.02.006

20. J. Jacod, Y. Li, P. A. Mykland, M. Podolskij, M. Vetter, Microstructure noise in the continuous case: the pre-averaging approach, *Stochastic Process. Appl.*, **119** (2009), 2249–2276. https://doi.org/10.1016/j.spa.2008.11.004

21. J. Jacod, M. Podolskij, M. Vetter, Limit theorems for moving averages of discretized processes plus noise, *Ann. Statist.*, **38** (2010), 1478–1545. https://doi.org/10.1214/09-AOS756

22. N. Hautsch, M. Podolskij, Preaveraging-based estimation of quadratic variation in the presence of noise and jumps: theory, implementation, and empirical evidence, *J. Bus. Econ. Stat.*, **31** (2013), 165–183. https://doi.org/10.1080/07350015.2012.754313

23. J. Jacod, Y. Li, X. Zheng, Statistical properties of microstructure noise, *Econometrica*, **85** (2017), 1133–1174. https://doi.org/10.3982/ECTA13085

24. Y. Aït-Sahalia, P. A. Mykland, L. Zhang, Ultra high frequency volatility estimation with dependent microstructure noise, *J. Econom.*, **160** (2011), 160–175. https://doi.org/10.1016/j.jeconom.2010.03.028
25. Z. M. Li, O. Linton, A ReMeDI for microstructure noise, *Econometrica*, **90** (2022), 367–389. https://doi.org/10.3982/ECTA17505
26. S. Peluso, A. Mira, P. Muliere, Conditionally Gaussian random sequences for an integrated variance estimator with correlation between noise and returns, *Appl. Stoch. Models Bus. Ind.*, **35** (2019), 1282–1297. https://doi.org/10.1002/asmb.2476
27. J. Jacod, Y. Li, X. Zheng, Estimating the integrated volatility with tick observations, *J. Econom.*, **208** (2019), 80–100. https://doi.org/10.1016/j.jeconom.2018.09.006
28. P. R. Hansen, A. Lunde, An unbiased measure of realized variance, *SSRN*, 2004, 524602.
29. Z. Liu, X. Kong, B. Jing, Estimating the integrated volatility using high-frequency data with zero durations, *J. Econom.*, **204** (2018), 18–32. https://doi.org/10.1016/j.jeconom.2017.12.008
30. Z. M. Li, O. Linton, Robust estimation of integrated and spot volatility, *J. Econom.*, **254** (2026), 105614. https://doi.org/10.1016/j.jeconom.2023.105614
31. W. Wasserfallen, H. Zimmermann, The behavior of intra-daily exchange rates, *J. Bank. Financ.*, **9** (1985), 55–72. https://doi.org/10.1016/0378-4266(85)90062-7
32. Z. M. Li, R. J. A. Laeven, M. H. Vellekoop, Dependent microstructure noise and integrated volatility estimation from high-frequency data, *J. Econom.*, **215** (2020), 536–558. https://doi.org/10.1016/j.jeconom.2019.10.004
33. B. Y. Jing, Z. Liu, X. B. Kong, On the estimation of integrated volatility with jumps and microstructure noise, *J. Bus. Econ. Stat.*, **32** (2014), 457–467. https://doi.org/10.1080/07350015.2014.906350
34. C. Brownlees, E. Nualart, Y. Sun, On the estimation of integrated volatility in the presence of jumps and microstructure noise, *Econom. Rev.*, **39** (2020), 991–1013. https://doi.org/10.1080/07474938.2020.1735751
35. M. Vetter, Estimation of integrated volatility of volatility with applications to goodness-of-fit testing, *Bernoulli*, **21** (2015), 2393–2418. https://doi.org/10.3150/14-BEJ648
36. Q. Han, Y. Liu, Estimation of serially correlated microstructure noises, *J. Quant. Tech. Econ.*, **4** (2007), 92–102.
37. P. Hall, C. C. Heyde, *Martingale limit theory and its application*, New York: Academic Press, 2014.
38. R. C. Bradley, *Introduction to strong mixing conditions*, Heber City: Kendrick Press, 2007.

Appendix

Appendix A. Proof of Lemma 2.1

By Assumption 2.2, we find that $\Delta\varepsilon_i$ is uncorrelated with $\varepsilon_{i\pm h}$ for any $h \geq k_n$, $k_n = \lceil n\theta_0 \rceil + 1$. So it follows that

$$0 = E\left[\Delta\varepsilon_i\left(\varepsilon_{i+k_n} - \varepsilon_{i-k_n-1}\right)\right] = E\left(\Delta\varepsilon_i \sum_{h=-k_n}^{k_n} \Delta\varepsilon_{i+h}\right) = \sum_{h=-k_n}^{k_n} \pi^{(h)}. \tag{A.1}$$

Again, ΔX_i is uncorrelated with $\varepsilon_{i\pm h}$ for any $h \geq k_n$. So we have

$$0 = E\left[\Delta X_i\left(\varepsilon_{i+k_n} - \varepsilon_{i-k_n-1}\right)\right] = E\left(\Delta X_i \sum_{h=-k_n}^{k_n} \Delta\varepsilon_{i+h}\right) = \sum_{h=-k_n}^{k_n} \gamma^{(h)}. \tag{A.2}$$

This completes the proof of Lemma 2.1. □

Appendix B. Proof of Theorem 3.1

As stated earlier, we condition on σ_t^2 in the following stochastic analysis. Thus, we treat σ_t^2 as a deterministic function in our derivations. We first prove the unbiasedness of the estimator $\mathrm{PKRV}(Y)^{(n)}$ for IV. Here, we calculate directly the expectation of $\mathrm{PKRV}(Y)^{(n)}$.

$$\begin{aligned}
E\left(\sum_{i=k_n}^{n-2k_n} \overline{Y}_i^2 \,\middle|\, \{\sigma_t^2\}_0^1\right) &= E\left\{\sum_{i=k_n}^{n-2k_n}\left[\sum_{j=1}^{k_n}\left(\frac{1}{\sqrt{k_n}}\Delta X_{i+j} + \frac{1}{\sqrt{k_n}}\Delta\varepsilon_{i+j}\right)\right]^2 \,\middle|\, \{\sigma_t^2\}_0^1\right\} \\
&= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\left(\frac{1}{k_n}\Delta X_{i+j}\Delta X_{i+m} + 2\frac{1}{k_n}\Delta X_{i+j}\Delta\varepsilon_{i+m} + \frac{1}{k_n}\Delta\varepsilon_{i+j}\Delta\varepsilon_{i+m}\right) \,\middle|\, \{\sigma_t^2\}_0^1\right] \\
&= (1) + (2) + (3).
\end{aligned} \tag{B.1}$$

Next, we derive the expressions of these three terms as follows:

$$\begin{aligned}
(1) &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta X_{i+j}\Delta X_{i+m}\right) \,\middle|\, \{\sigma_t^2\}_0^1\right] \\
&= \sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}E\left[\left(\Delta X_{i+j}\Delta X_{i+m}\right) \,\middle|\, \{\sigma_t^2\}_0^1\right] \\
&= \int_0^1 \sigma_t^2\, dt + O_p\left(\frac{k_n}{n}\right).
\end{aligned}$$

By Assumptions 2.1 and 2.2, (2) and (3) are given by

$$(2) = 2E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta X_{i+j}\Delta\varepsilon_{i+m}\right) \,\middle|\, \{\sigma_t^2\}_0^1\right] = \frac{2}{k_n}\sum_{i=k_n}^{n-2k_n}\left[k_n\gamma^{(0)} + \sum_{j=1}^{k_n}(k_n - j)\left(\gamma^{(j)} + \gamma^{(-j)}\right)\right].$$

$$(3) = E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta\varepsilon_{i+j}\Delta\varepsilon_{i+m}\right) \,\middle|\, \{\sigma_t^2\}_0^1\right] = \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left[k_n\pi^{(0)} + \sum_{j=1}^{k_n}(k_n - j)\left(\pi^{(j)} + \pi^{(-j)}\right)\right].$$

$$\begin{aligned}
E\left[\sum_{i=k_n}^{n-2k_n}(\overline{Y}_i\overline{Y}_{i+k_n}) \,\middle|\, \{\sigma_t^2\}_0^1\right] &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta Y_{i+j}\Delta Y_{i+k_n+m}\right) \,\middle|\, \{\sigma_t^2\}_0^1\right] \\
&= E\left[\sum_{i=k_n}^{n-2k_n}\frac{1}{k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\left(\Delta X_{i+j}\Delta X_{i+k_n+m} + \Delta X_{i+j}\Delta\varepsilon_{i+k_n+m}\right.\right. \\
&\qquad \left.\left. + \Delta\varepsilon_{i+j}\Delta X_{i+k_n+m} + \Delta\varepsilon_{i+j}\Delta\varepsilon_{i+k_n+m}\right) \,\middle|\, \{\sigma_t^2\}_0^1\right] \\
&= (1) + (2) + (3) + (4).
\end{aligned} \tag{B.2}$$

Since the efficient price is a Brownian semimartingale process, ΔX_{i+j} and ΔX_{i+k_n+m} are uncorrelated for any $j(m) \in [1, k_n]$. So (1) is given by

$$
\begin{aligned}
(1) &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta X_{i+j}\Delta X_{i+k_n+m}\right)\middle|\,\{\sigma_t^2\}_0^1\right]\\
&= \sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}E\left[\left(\Delta X_{i+j}\Delta X_{i+k_n+m}\right)\,\middle|\,\{\sigma_t^2\}_0^1\right]\\
&= 0.
\end{aligned}
$$

By Assumption 2.1 and 2.2, we have

$$
\begin{aligned}
(2) &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta X_{i+j}\Delta \varepsilon_{i+k_n+m}\right)\middle|\,\{\sigma_t^2\}_0^1\right]\\
&= \sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\gamma^{(k_n+m-j)}\\
&= \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\gamma^{(j)}.
\end{aligned}
$$

$$
\begin{aligned}
(3) &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta \varepsilon_{i+j}\Delta X_{i+k_n+m}\right)\middle|\,\{\sigma_t^2\}_0^1\right]\\
&= \sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\gamma^{(j-k_n-m)}\\
&= \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\gamma^{(-j)}.
\end{aligned}
$$

$$
\begin{aligned}
(4) &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta \varepsilon_{i+j}\Delta \varepsilon_{i+k_n+m}\right)\middle|\,\{\sigma_t^2\}_0^1\right]\\
&= \sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\pi^{(k_n+m-j)}\\
&= \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\pi^{(j)}.
\end{aligned}
$$

By using the same methods used in Eq (B.2), we have

$$
\begin{aligned}
E\left[\sum_{i=k_n}^{n-2k_n}(\overline{Y}_i\overline{Y}_{i-k_n})\,\middle|\,\{\sigma_t^2\}_0^1\right] &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta Y_{i+j}\Delta Y_{i-k_n+m}\right)\,\middle|\,\{\sigma_t^2\}_0^1\right] \\
&= E\left[\sum_{i=k_n}^{n-2k_n}\frac{1}{k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\left(\Delta X_{i+j}\Delta X_{i-k_n+m} + \Delta X_{i+j}\Delta\varepsilon_{i-k_n+m}\right.\right. \\
&\qquad \left.\left. +\Delta\varepsilon_{i+j}\Delta X_{i-k_n+m} + \Delta\varepsilon_{i+j}\Delta\varepsilon_{i-k_n+m}\right)\,\middle|\,\{\sigma_t^2\}_0^1\right] \\
&= (1) + (2) + (3) + (4).
\end{aligned}
\tag{B.3}
$$

$$
\begin{aligned}
(1) &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta X_{i+j}\Delta X_{i-k_n+m}\right)\,\middle|\,\{\sigma_t^2\}_0^1\right] \\
&= \sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}E\left[\left(\Delta X_{i+j}\Delta X_{i-k_n+m}\right)\,\middle|\,\{\sigma_t^2\}_0^1\right] \\
&= 0. \\
(2) &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta X_{i+j}\Delta\varepsilon_{i-k_n+m}\right)\,\middle|\,\{\sigma_t^2\}_0^1\right] \\
&= \sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\gamma^{(m-j-k_n)} \\
&= \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\gamma^{(-j)}. \\
(3) &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta\varepsilon_{i+j}\Delta X_{i-k_n+m}\right)\,\middle|\,\{\sigma_t^2\}_0^1\right] \\
&= \sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\gamma^{(j-m+k_n)} \\
&= \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\gamma^{(j)}. \\
(4) &= E\left[\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\left(\Delta\varepsilon_{i+j}\Delta\varepsilon_{i-k_n+m}\right)\,\middle|\,\{\sigma_t^2\}_0^1\right] \\
&= \sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n}\sum_{m=1}^{k_n}\frac{1}{k_n}\pi^{(m-j-k_n)} \\
&= \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\pi^{(-j)}.
\end{aligned}
$$

By Lemma 2.1, we have

$$
\begin{aligned}
E\left[\mathrm{PKRV}(Y)^{(n)} \,\middle|\, \{\sigma_t^2\}_0^1\right] &= E\left[\sum_{i=k_n}^{n-2k_n}\left(\overline{Y}_i^2 + \overline{Y}_i\overline{Y}_{i+k_n} + \overline{Y}_i\overline{Y}_{i-k_n}\right) \,\middle|\, \{\sigma_t^2\}_0^1\right] \\
&= \int_0^1 \sigma_t^2\,dt + \frac{2}{k_n}\sum_{i=k_n}^{n-2k_n}\left[k_n\gamma^{(0)} + \sum_{j=1}^{k_n}(k_n - j)(\gamma^{(j)} + \gamma^{(-j)})\right] \\
&\quad + \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left[k_n\pi^{(0)} + \sum_{j=1}^{k_n}(k_n - j)(\pi^{(j)} + \pi^{(-j)})\right] \\
&\quad + \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\gamma^{(j)} + \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\gamma^{(-j)} + \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\pi^{(j)} \\
&\quad + \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\gamma^{(-j)} + \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\gamma^{(j)} + \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{j=1}^{k_n} j\pi^{(-j)} + O\left(\frac{k_n}{n}\right) \\
&= \int_0^1 \sigma_t^2\,dt + 2\sum_{i=k_n}^{n-2k_n}\left[\gamma^{(0)} + \sum_{j=1}^{k_n}(\gamma^{(j)} + \gamma^{(-j)})\right] + \sum_{i=k_n}^{n-2k_n}\left[\pi^{(0)} + \sum_{j=1}^{k_n}(\pi^{(j)} + \pi^{(-j)})\right] + O\left(\frac{k_n}{n}\right) \\
&= \int_0^1 \sigma_t^2\,dt + \sum_{i=k_n}^{n-2k_n}\left[2\sum_{j=-k_n}^{k_n}\gamma^{(j)} + \sum_{j=-k_n}^{k_n}\pi^{(j)}\right] + O\left(\frac{k_n}{n}\right) \\
&= \int_0^1 \sigma_t^2\,dt + O_p\left(\frac{k_n}{n}\right).
\end{aligned}
\tag{B.4}
$$

The proof of the unbiasedness of the estimator $\mathrm{PKRV}(Y)^{(n)}$ for IV is finished.

Now we are left to prove the central limit theorems of the estimator $\mathrm{PKRV}(Y)^{(n)}$ for IV. We first rewrite the $\mathrm{PKRV}(Y)^{(n)}$ as follows:

$$
\begin{aligned}
\mathrm{PKRV}(Y)^{(n)} &= \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left(X_{i+k_n} - X_i + \varepsilon_{i+k_n} - \varepsilon_i\right)\left(X_{i+2k_n} - X_{i-k_n} + \varepsilon_{i+2k_n} - \varepsilon_{i-k_n}\right) \\
&= \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left(X_{i+k_n} - X_i\right)\left(X_{i+2k_n} - X_{i-k_n}\right) + \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left(X_{i+k_n} - X_i\right)\left(\varepsilon_{i+2k_n} - \varepsilon_{i-k_n}\right) \\
&+ \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left(\varepsilon_{i+k_n} - \varepsilon_i\right)\left(X_{i+2k_n} - X_{i-k_n}\right) + \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left(\varepsilon_{i+k_n} - \varepsilon_i\right)\left(\varepsilon_{i+2k_n} - \varepsilon_{i-k_n}\right) \\
&= R + U + V + W,
\end{aligned}
\tag{B.5}
$$

where

$$
R = \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left(X_{i+k_n} - X_i\right)\left(X_{i+2k_n} - X_{i-k_n}\right), \quad U = \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left(X_{i+k_n} - X_i\right)\left(\varepsilon_{i+2k_n} - \varepsilon_{i-k_n}\right),
$$

$$
V = \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left(\varepsilon_{i+k_n} - \varepsilon_i\right)\left(X_{i+2k_n} - X_{i-k_n}\right), \quad W = \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\left(\varepsilon_{i+k_n} - \varepsilon_i\right)\left(\varepsilon_{i+2k_n} - \varepsilon_{i-k_n}\right).
$$

The variances of R, U, V, and W are given by

$$\begin{aligned}\operatorname{var}\left(R \mid \{\sigma_t^2\}_0^1\right) &= \frac{1}{k_n^2}\sum_{i=k_n}^{n-2k_n} \operatorname{var}\left\{\left[(X_{i+k_n}-X_i)(X_{i+2k_n}-X_{i-k_n})\right] \mid \{\sigma_t^2\}_0^1\right\} \\ &\quad + \frac{1}{k_n^2}\sum_{\substack{i,j=k_n\\ i\neq j}}^{n-2k_n} \operatorname{cov}\left\{\left[(X_{i+k_n}-X_i)(X_{i+2k_n}-X_{i-k_n}),\left(X_{j+k_n}-X_j\right)\left(X_{j+2k_n}-X_{j-k_n}\right)\right] \mid \{\sigma_t^2\}_0^1\right\} \\ &= (1)+(2).\end{aligned} \tag{B.6}$$

Suppose that the efficient price obeys a Brownian semimartingale process, (1) and (2) above are given by

$$\begin{aligned}(1) &= \frac{1}{k_n^2}\sum_{i=k_n}^{n-2k_n}\left\{E\left[(X_{i+k_n}-X_i)(X_{i+2k_n}-X_{i-k_n}) \mid \{\sigma_t^2\}_0^1\right]^2 - \left(E\left[(X_{i+k_n}-X_i)(X_{i+2k_n}-X_{i-k_n}) \mid \{\sigma_t^2\}_0^1\right]\right)^2\right\} \\ &= 4\sum_{i=k_n}^{n-2k_n}\sigma_i^4\frac{1}{n^2}, \\ (2) &= \frac{1}{k_n^2}\sum_{\substack{i,j=k_n\\ i\neq j}}^{n-2k_n}\Bigg\{E\left[(X_{i+k_n}-X_i)(X_{i+2k_n}-X_{i-k_n})(X_{j+k_n}-X_j)(X_{j+2k_n}-X_{j-k_n}) \mid \{\sigma_t^2\}_0^1\right] \\ &\quad - E\left[(X_{i+k_n}-X_i)(X_{i+2k_n}-X_{i-k_n}) \mid \{\sigma_t^2\}_0^1\right]\cdot E\left[(X_{j+k_n}-X_j)(X_{j+2k_n}-X_{j-k_n}) \mid \{\sigma_t^2\}_0^1\right]\Bigg\} \\ &= \frac{1}{k_n^2}\sum_{\substack{i,j=k_n\\ i\neq j}}^{n-2k_n}\Bigg\{E\left[(X_{i+k_n}-X_i)^2(X_{j+k_n}-X_j)^2 \mid \{\sigma_t^2\}_0^1\right] \\ &\quad + E\left[(X_{i+k_n}-X_i)^2(X_{j+k_n}-X_j)(X_{j+2k_n}-X_{j+k_n}) \mid \{\sigma_t^2\}_0^1\right] \\ &\quad + E\left[(X_{i+k_n}-X_i)^2(X_{j+k_n}-X_j)(X_j-X_{j-k_n}) \mid \{\sigma_t^2\}_0^1\right] \\ &\quad + E\left[(X_{i+k_n}-X_i)(X_{i+2k_n}-X_{i+k_n})(X_{j+k_n}-X_j)^2 \mid \{\sigma_t^2\}_0^1\right] \\ &\quad + E\left[(X_{i+k_n}-X_i)(X_i-X_{i-k_n})(X_{j+k_n}-X_j)^2 \mid \{\sigma_t^2\}_0^1\right] \\ &\quad + E\left[(X_{i+k_n}-X_i)(X_{i+2k_n}-X_{i+k_n})(X_{j+k_n}-X_j)(X_{j+2k_n}-X_{j+k_n}) \mid \{\sigma_t^2\}_0^1\right] \\ &\quad + E\left[(X_{i+k_n}-X_i)(X_{i+2k_n}-X_{i+k_n})(X_{j+k_n}-X_j)(X_j-X_{j-k_n}) \mid \{\sigma_t^2\}_0^1\right] \\ &\quad + E\left[(X_{i+k_n}-X_i)(X_i-X_{i-k_n})(X_{j+k_n}-X_j)(X_{j+2k_n}-X_{j+k_n}) \mid \{\sigma_t^2\}_0^1\right] \\ &\quad + E\left[(X_{i+k_n}-X_i)(X_i-X_{i-k_n})(X_{j+k_n}-X_j)(X_j-X_{j-k_n}) \mid \{\sigma_t^2\}_0^1\right] \\ &\quad - E\left[(X_{i+k_n}-X_i)(X_{i+2k_n}-X_{i-k_n}) \mid \{\sigma_t^2\}_0^1\right]\cdot E\left[(X_{j+k_n}-X_j)(X_{j+2k_n}-X_{j-k_n}) \mid \{\sigma_t^2\}_0^1\right]\Bigg\} \\ &= \frac{1}{k_n^2}\sum_{i=k_n}^{n-2k_n}\left\{\sum_{m=1}^{k_n}\left[10\sigma_i^4\frac{(k_n-m)^2}{n^2}+12\sigma_i^4\frac{(k_n-m)m}{n^2}+4\sigma_i^4\frac{m^2}{n^2}\right]-2\sigma_i^4\frac{k_n^3}{n^2}\right\}\end{aligned}$$

$$= \sum_{i=k_n}^{n-2k_n} \sigma_i^4 \left(\frac{14k_n}{3n^2} - \frac{3}{n^2} + \frac{1}{3n^2 k_n} \right).$$

Then

$$\operatorname{var}(R|\{\sigma_t^2\}_0^1) = \sum_{i=k_n}^{n-2k_n} \sigma_i^4 \left(\frac{14k_n}{3n^2} + \frac{1}{n^2} + \frac{1}{3n^2 k_n} \right), \tag{B.7}$$

$$\begin{aligned}
\operatorname{var}\left(R \mid \{\sigma_t^2\}_0^1\right) &= \frac{1}{k_n^2} \sum_{i=k_n}^{n-2k_n} \operatorname{var}\left\{ [(X_{i+k_n} - X_i)(\varepsilon_{i+2k_n} - \varepsilon_{i-k_n})] \mid \{\sigma_t^2\}_0^1 \right\} \\
&\quad + \frac{1}{k_n^2} \sum_{\substack{i,j=k_n \\ i \neq j}}^{n-2k_n} \operatorname{cov}\left\{ \left[(X_{i+k_n} - X_i)(\varepsilon_{i+2k_n} - \varepsilon_{i-k_n}), (X_{j+k_n} - X_j)(\varepsilon_{j+2k_n} - \varepsilon_{j-k_n}) \right] \mid \{\sigma_t^2\}_0^1 \right\} \\
&= (1) + (2).
\end{aligned}$$

If we assume that Assumption 2.2 holds and let $\mathrm{E}\,\varepsilon_i^2 = \omega^2$, (1) above is given by

$$\begin{aligned}
(1) &= \frac{1}{k_n^2} \sum_{i=k_n}^{n-2k_n} \operatorname{var}\left\{ [(X_{i+k_n} - X_i)(\varepsilon_{i+2k_n} - \varepsilon_{i-k_n})] \mid \{\sigma_t^2\}_0^1 \right\} \\
&= \frac{1}{k_n^2} \sum_{i=k_n}^{n-2k_n} \mathrm{E}\left\{ [(X_{i+k_n} - X_i)(\varepsilon_{i+2k_n} - \varepsilon_{i-k_n})]^2 \mid \{\sigma_t^2\}_0^1 \right\} \\
&= \frac{1}{k_n^2} \sum_{i=k_n}^{n-2k_n} \mathrm{E}\left\{ (X_{i+k_n} - X_i)^2 \left(\varepsilon_{i+2k_n}^2 - 2\varepsilon_{i+2k_n}\varepsilon_{i-k_n} + \varepsilon_{i-k_n}^2 \right) \mid \{\sigma_t^2\}_0^1 \right\} \\
&= \frac{1}{k_n} \sum_{i=k_n}^{n-2k_n} 2\sigma_i^2 \omega^2 \frac{1}{n.}
\end{aligned}$$

If we let $A_i = (X_{i+k_n} - X_i)(\varepsilon_{i+2k_n} - \varepsilon_{i-k_n})$ and let j be such that $|j - i| \geq 4k_n$, it can be easily proved that $\operatorname{cov}(A_i, A_j) = 0$. Without losing of generality, we can set

$$\operatorname{cov}[(A_i, A_j) \mid \{\sigma_t^2\}_0^1] = a\rho^{|j-i|} \operatorname{var}[A_i \mid \{\sigma_t^2\}_0^1], \quad \rho \in (0, 1), \text{ and } a \in [-1, 1].$$

For a proof of this setting, see Theorem A.5 in [37]; see also [38]. Throughout the proof, a and ρ denotes a generic constant[†] and $\rho \in (0, 1)$. So we have

$$\begin{aligned}
(2) &= \frac{1}{k_n^2} \sum_{\substack{i,j=k_n \\ i \neq j}}^{n-2k_n} \operatorname{cov}\left[(A_i, A_j) \mid \{\sigma_t^2\}_0^1 \right] = \frac{2}{k_n^2} \sum_{\substack{i,j=k_n \\ j > i}}^{n-2k_n} a\rho^{|j-i|} \operatorname{var}\left[A_i \mid \{\sigma_t^2\}_0^1 \right] \\
&= \frac{1}{k_n} \sum_{i=k_n}^{n-2k_n} \sum_{m=1}^{3k_n} 4a\rho^m \sigma_i^2 \omega^2 \frac{1}{n} = \frac{1}{k_n} \sum_{i=k_n}^{n-2k_n} 4\sigma_i^2 \omega^2 \frac{1}{n} \frac{a\rho}{1-\rho}.
\end{aligned}$$

[†] In this paper, for simplicity, we adopt generic constants, e.g. ρ, γ and K, so the simple notations do not affect our conclusions.

By adding up (1) and (2), we have

$$\begin{aligned}\operatorname{var}\left(U \mid \{\sigma_t^2\}_0^1\right) &= \sum_{i=k_n}^{n-2k_n}\left(2\sigma_i^2\omega^2\frac{1}{nk_n} + 4\sigma_i^2\omega^2\frac{a\rho}{1-\rho}\frac{1}{nk_n}\right)\\ &= \frac{2\omega^2}{k_n}\frac{1+(2a-1)\rho}{1-\rho}\sum_{i=k_n}^{n-2k_n}\sigma_i^2\frac{1}{n}\\ &= \frac{2\omega^2}{k_n}\frac{1+(2a-1)\rho}{1-\rho}\int_0^1\sigma_t^2\,dt + O_p\left(\frac{1}{n}\right).\end{aligned}\tag{B.8}$$

Next, we mainly use the techniques used in B.8. We only state the proof of key steps in the calculation of var(V) and var(W).

$$\begin{aligned}\operatorname{var}\left(V \mid \{\sigma_t^2\}_0^1\right) &= \frac{1}{k_n^2}\sum_{i=k_n}^{n-2k_n}\operatorname{var}\left\{[(\varepsilon_{i+k_n}-\varepsilon_i)(X_{i+2k_n}-X_{i-k_n})] \mid \{\sigma_t^2\}_0^1\right\}\\ &\quad+\frac{1}{k_n^2}\sum_{\substack{i,j=k_n\\ i\neq j}}^{n-2k_n}\operatorname{cov}\left\{\left[(\varepsilon_{i+k_n}-\varepsilon_i)(X_{i+2k_n}-X_{i-k_n}),(\varepsilon_{j+k_n}-\varepsilon_j)(X_{j+2k_n}-X_{j-k_n})\right] \mid \{\sigma_t^2\}_0^1\right\}\\ &= (1)+(2),\end{aligned}$$

$$(1) = \frac{1}{k_n^2}\sum_{i=k_n}^{n-2k_n}\operatorname{var}\left\{[(\varepsilon_{i+k_n}-\varepsilon_i)(X_{i+2k_n}-X_{i-k_n})] \mid \{\sigma_t^2\}_0^1\right\} = \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}6\sigma_i^2\omega^2\frac{1}{n},$$

$$\begin{aligned}(2) &= \frac{1}{k_n^2}\sum_{\substack{i,j=k_n\\ i\neq j}}^{n-2k_n}\operatorname{cov}\left[(A_i,A_j) \mid \{\sigma_t^2\}_0^1\right] = \frac{2}{k_n^2}\sum_{\substack{i,j=k_n\\ j>i}}^{n-2k_n}a\rho^{|j-i|}\operatorname{var}\left[A_i \mid \{\sigma_t^2\}_0^1\right]\\ &= \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}\sum_{m=1}^{3k_n}12a\rho^m\sigma_i^2\omega^2\frac{1}{n} = \frac{1}{k_n}\sum_{i=k_n}^{n-2k_n}12a\sigma_i^2\omega^2\frac{1}{n}\frac{\rho}{1-\rho}.\end{aligned}$$

We then get

$$\begin{aligned}\operatorname{var}\left(V \mid \{\sigma_t^2\}_0^1\right) &= \sum_{i=k_n}^{n-2k_n}\left(6\sigma_i^2\omega^2\frac{1}{nk_n} + 12a\sigma_i^2\omega^2\frac{\rho}{1-\rho}\frac{1}{nk_n}\right) = \frac{6\omega^2}{k_n}\frac{1+(2a-1)\rho}{1-\rho}\sum_{i=k_n}^{n-2k_n}\sigma_i^2\frac{1}{n}\\ &= \frac{6\omega^2}{k_n}\frac{1+(2a-1)\rho}{1-\rho}\int_0^1\sigma_t^2\,dt + O_p\left(\frac{1}{n}\right),\end{aligned}\tag{B.9}$$

$$\begin{aligned}\operatorname{var}\left(W \mid \{\sigma_t^2\}_0^1\right) &= \frac{1}{k_n^2}\sum_{i=k_n}^{n-2k_n}\operatorname{var}\left\{[(\varepsilon_{i+k_n}-\varepsilon_i)(\varepsilon_{i+2k_n}-\varepsilon_{i-k_n})] \mid \{\sigma_t^2\}_0^1\right\}\\ &\quad+\frac{1}{k_n^2}\sum_{\substack{i,j=k_n\\ i\neq j}}^{n-2k_n}\operatorname{cov}\left\{\left[(\varepsilon_{i+k_n}-\varepsilon_i)(\varepsilon_{i+2k_n}-\varepsilon_{i-k_n}),(\varepsilon_{j+k_n}-\varepsilon_j)(\varepsilon_{j+2k_n}-\varepsilon_{j-k_n})\right] \mid \{\sigma_t^2\}_0^1\right\}\\ &= (1)+(2),\end{aligned}$$

where

$$
\begin{aligned}
(1) &= \frac{1}{k_n^2} \sum_{i=k_n}^{n-2k_n} \operatorname{var}\left\{ [(\varepsilon_{i+k_n} - \varepsilon_i)(\varepsilon_{i+2k_n} - \varepsilon_{i-k_n})] \;\middle|\; \{\sigma_t^2\}_0^1 \right\} \\
&= \frac{4}{k_n^2} \sum_{i=k_n}^{n-2k_n} \omega^4, \\
(2) &= \frac{1}{k_n^2} \sum_{\substack{i,j=k_n \\ i \neq j}}^{n-2k_n} \operatorname{cov}\left[(A_i, A_j) \;\middle|\; \{\sigma_t^2\}_0^1 \right] \\
&= \frac{2}{k_n^2} \sum_{\substack{i,j=k_n \\ j>i}}^{n-2k_n} a\rho^{|j-i|} \operatorname{var}\left[A_i \;\middle|\; \{\sigma_t^2\}_0^1 \right] \\
&= \frac{1}{k_n^2} \sum_{i=k_n}^{n-2k_n} 8a\omega^4 \frac{\rho}{1-\rho}.
\end{aligned}
$$

Then

$$
\begin{aligned}
\operatorname{var}\left(W \;\middle|\; \{\sigma_t^2\}_0^1 \right) &= \sum_{i=k_n}^{n-2k_n} \left(\frac{4\omega^4}{k_n^2} + \frac{8a\omega^4}{k_n^2} \frac{\rho}{1-\rho} \right) \\
&= \frac{4[1+(2a-1)\rho]}{(1-\rho)k_n^2} \sum_{i=k_n}^{n-2k_n} \omega^4 \\
&= \frac{4[1+(2a-1)\rho]}{(1-\rho)k_n^2} n\omega^4 + O_p\left(\frac{1}{k_n} \right).
\end{aligned} \tag{B.10}
$$

By the definition of the correlation coefficient, we have

$$
\begin{aligned}
\operatorname{cov}\left[(R, U) \;\middle|\; \{\sigma_t^2\}_0^1 \right] &= \gamma \left[\frac{28\omega^2[1+(2a-1)\rho]}{3(1-\rho)} \int_0^1 \sigma_t^4 \, dt \int_0^1 \sigma_t^2 \, dt \right]^{1/2} n^{-1/2} = O_p(n^{-1/2}), \\
\operatorname{cov}\left[(R, V) \;\middle|\; \{\sigma_t^2\}_0^1 \right] &= \gamma \left[\frac{84\omega^2[1+(2a-1)\rho]}{3(1-\rho)} \int_0^1 \sigma_t^4 \, dt \int_0^1 \sigma_t^2 \, dt \right]^{1/2} n^{-1/2} = O_p(n^{-1/2}), \\
\operatorname{cov}\left[(U, V) \;\middle|\; \{\sigma_t^2\}_0^1 \right] &= \gamma 2\sqrt{3} \left[\frac{[1+(2a-1)\rho]\omega^2}{(1-\rho)} \int_0^1 \sigma_t^2 \, dt \right] k_n^{-1} = O_p(k_n^{-1}), \\
\operatorname{cov}\left[(U, W) \;\middle|\; \{\sigma_t^2\}_0^1 \right] &= \gamma 2\sqrt{2} \, \frac{[1+(2a-1)\rho]}{(1-\rho)} \omega^3 \left(\int_0^1 \sigma_t^2 \, dt \right)^{1/2} \left(\frac{n}{k_n^3} \right)^{1/2} = O_p(n^{1/2} k_n^{-3/2}), \\
\operatorname{cov}\left[(V, W) \;\middle|\; \{\sigma_t^2\}_0^1 \right] &= \gamma 2\sqrt{6} \, \frac{[1+(2a-1)\rho]}{(1-\rho)} \omega^3 \left(\int_0^1 \sigma_t^2 \, dt \right)^{1/2} \left(\frac{n}{k_n^3} \right)^{1/2} = O_p(n^{1/2} k_n^{-3/2}), \\
\operatorname{cov}\left[(R, W) \;\middle|\; \{\sigma_t^2\}_0^1 \right] &= \gamma \left[\frac{56[1+(2a-1)\rho]}{3(1-\rho)} \omega^4 \int_0^1 \sigma_t^4 \, dt \right]^{1/2} k_n^{-1/2} = O_p(k_n^{-1/2}),
\end{aligned}
$$

where $\gamma \in (-1, 1)$ are correlation coefficients, and γ denotes a generic constant.

By adding up the variances of R, U, V, W, and their covariances, we find

$$\begin{aligned}
\operatorname{var}\Big(\mathrm{PKRV}(Y)^n \;\Big|\; \{\sigma_t^2\}_0^1\Big) &= \operatorname{var}\Big(R \;\Big|\; \{\sigma_t^2\}_0^1\Big) + \operatorname{var}\Big(U \;\Big|\; \{\sigma_t^2\}_0^1\Big) + \operatorname{var}\Big(V \;\Big|\; \{\sigma_t^2\}_0^1\Big) + \operatorname{var}\Big(W \;\Big|\; \{\sigma_t^2\}_0^1\Big) \\
&\quad + 2\operatorname{cov}\Big[(R,U) \;\Big|\; \{\sigma_t^2\}_0^1\Big] + 2\operatorname{cov}\Big[(R,V) \;\Big|\; \{\sigma_t^2\}_0^1\Big] + 2\operatorname{cov}\Big[(R,W) \;\Big|\; \{\sigma_t^2\}_0^1\Big] \\
&\quad + 2\operatorname{cov}\Big[(U,V) \;\Big|\; \{\sigma_t^2\}_0^1\Big] + 2\operatorname{cov}\Big[(U,W) \;\Big|\; \{\sigma_t^2\}_0^1\Big] + 2\operatorname{cov}\Big[(V,W) \;\Big|\; \{\sigma_t^2\}_0^1\Big] \\
&= \frac{14k_n}{3n}\int_0^1 \sigma_t^4\,dt + \frac{4[1+(2a-1)\rho]\omega^4 n}{(1-\rho)k_n^2} \\
&\quad + \frac{\gamma}{k_n^{1/2}}\left[\frac{56[1+(2a-1)\rho]}{3(1-\rho)}\omega^4\int_0^1 \sigma_t^4\,dt\right]^{1/2} + O_p\Big(n^{-1/2}\Big).
\end{aligned}$$

To reduce the MSE of $\mathrm{PKRV}(Y)^{(n)}$, we set $\frac{\partial MSE}{\partial k_n} = 0$, and thus the optimal window width k_n^* is given by

$$k_n^* = \left[\frac{(168)^{1/2}\gamma + \Big(168\gamma^2 + 5376\Big)^{1/2}}{56}\right]^{2/3}\left[\frac{[1+(2a-1)\rho]\omega^4}{(1-\rho)\int_0^1 \sigma_t^4\,dt}\right]^{1/3} n^{2/3}.$$

If we take $k_n = k_n^*$, then,

$$\operatorname{var}\Big(\mathrm{PKRV}^{(n)} \;\Big|\; \{\sigma_t^2\}_0^1\Big) = \Gamma\, n^{-1/3}, \tag{B.11}$$

where Γ is the asymptotic conditional variance. The detailed expression of Γ is obvious but sophisticated. We complete the proof of Theorem 3.1. □

Appendix C. Proof of Lemma 3.1

Suppose that the jump process J_t is independent of the continuous efficient price process X_t and the microstructure ε_t, and suppose that the jump size k_{τ_i} is also i.i.d. and independent of the Poisson process N_t, we have

$$\mathrm{E}\left(\sum_{i=k_n}^{n-2k_n}\overline{Z}_i^2\right) = \mathrm{E}\left[\sum_{i=k_n}^{n-2k_n}\Big(\overline{Y}_i^2 + 2\overline{Y}_i\overline{J}_i + \overline{J}_i^2\Big)\right] = \mathrm{E}\left(\sum_{i=k_n}^{n-2k_n}\overline{Y}_i^2\right) + \mathrm{E}\left(\sum_{i=k_n}^{n-2k_n}k_{\tau_i}^2\right),$$

$$\begin{aligned}
&\mathrm{E}\Big[\sum_{i=k_n}^{n-2k_n}(\overline{Z}_i\overline{Z}_{i+k_n} + \overline{Z}_i\overline{Z}_{i-k_n})\Big] \\
&= \mathrm{E}\Big[\sum_{i=k_n}^{n-2k_n}(\overline{Y}_i\overline{Y}_{i+k_n} + \overline{Y}_i\overline{Y}_{i-k_n})\Big] \\
&\quad + \mathrm{E}\Big[\sum_{i=k_n}^{n-2k_n}(\overline{Y}_i\overline{J}_{i+k_n} + \overline{J}_i\overline{Y}_{i+k_n} + \overline{J}_i\overline{J}_{i+k_n} + \overline{Y}_i\overline{J}_{i-k_n} + \overline{J}_i\overline{Y}_{i-k_n} + \overline{J}_i\overline{J}_{i-k_n})\Big] \\
&= \mathrm{E}\Big[\sum_{i=k_n}^{n-2k_n}(\overline{Y}_i\overline{Y}_{i+k_n} + \overline{Y}_i\overline{Y}_{i-k_n})\Big].
\end{aligned}$$

This completes the proof of Lemma 3.1. □

Appendix D. Proof of Lemma 3.2

Following the framework of the proof of Theorem 1 in [33], we introduce a variable

$$\begin{aligned}
\eta_i^n &= |\overline{Z}_i|^2 1_{\{|\overline{Z}_i|\le\Theta(\Delta_n)\}} + \overline{Z}_i\overline{Z}_{i+k_n} + \overline{Z}_i\overline{Z}_{i-k_n} - |\overline{Y}_i|^2 - \overline{Y}_i\overline{Y}_{i+k_n} - \overline{Y}_i\overline{Y}_{i-k_n} \\
&= |\overline{Y}_i + \overline{J}_i|^2 1_{\{|\overline{Z}_i|\le\Theta(\Delta_n)\}} - |\overline{Y}_i|^2 + \overline{Y}_i\overline{J}_{i+k_n} + \overline{J}_i\overline{Y}_{i+k_n} + \overline{J}_i\overline{J}_{i+k_n} + \overline{Y}_i\overline{J}_{i-k_n} + \overline{J}_i\overline{Y}_{i-k_n} + \overline{J}_i\overline{J}_{i-k_n} \\
&= \eta_{i1}^n + \eta_{i2}^n + \eta_{i3}^n,
\end{aligned}$$

where

$$\begin{aligned}
\eta_{i1}^n &= |\overline{Y}_i + \overline{J}_i|^2\, 1_{\{|\overline{Z}_i|\le\Theta(\Delta_n)\}} - |\overline{Y}_i|^2, \\
\eta_{i2}^n &= \overline{Y}_i\, \overline{J}_{i+k_n} + \overline{J}_i\, \overline{Y}_{i+k_n} + \overline{J}_i\, \overline{J}_{i+k_n}, \\
\eta_{i3}^n &= \overline{Y}_i\, \overline{J}_{i-k_n} + \overline{J}_i\, \overline{Y}_{i-k_n} + \overline{J}_i\, \overline{J}_{i-k_n}.
\end{aligned}$$

The following two elementary inequalities will be used in the proof

$$\begin{aligned}
&\left||X+Y|^P - |X|^p\right| \le |Y|^p, \quad \text{when } p \le 1, \\
&\left||X+Y|^P - |X|^p\right| \le K\left(|Y|^p + |X|^{p-1}|Y|\right), \quad \text{When } p > 1,
\end{aligned}$$

Throughout the proof, K denotes a generic constant. When $|\overline{Z}_i| > \Theta(\Delta_n)$, we have

$$\left|\eta_{i1}^n\right| = |\overline{Y}_i|^2\, 1_{\{|\overline{Z}_i|>\Theta(\Delta_n)\}}.$$

When $|\overline{Z}_i| \le \Theta(\Delta_n)$, for an appropriate constant K, we deduce that

$$\left|\eta_{i1}^n\right| \le K\left[\left(|\overline{J}_i| \wedge \Theta(\Delta_n)\right)^2 + |\overline{Y}_i|\, |\overline{J}_i| \wedge \Theta(\Delta_n)\right].$$

In view of the boundedness assumption of the parameters, and by using the Hölder's and Burkholder's inequalities, we have

$$\begin{aligned}
&\mathrm{E}\left(|\overline{J}_i|^2\right) = \int_{i\Delta_n}^{(i+k_n)\Delta_n} \int_R \frac{1}{k_n} x^2 F_t(dx)\, ds \le K\Delta_n, \\
&\mathrm{E}\left(|\overline{X}_i|^p\right) \le K\Delta_n^{p/2}, \\
&\mathrm{E}\left(|\overline{\varepsilon}_i|^p\right) \le K/k_n^{p/2} = K\Delta_n^{p/3},
\end{aligned}$$

where $F_t(dx)$ is a Poisson random measure. For any $p > 0$, by applying Minkowski's inequality to the process Y, we have

$$\mathrm{E}\left(\left|\overline{Y}_i\right|^p\right) \le K\Delta_n^{p/2} + K\Delta_n^{p/3} = K\Delta_n^{p/3}, \quad \text{to first order.}$$

By applying Eq (6.25) of Jacod (2012) to the process J, we get

$$\mathrm{E}\left(|\overline{J}_i| \wedge \Theta(\Delta_n)\right)^2 \le K\Delta_n\Theta(\Delta_n)^{2-s}, \quad \text{for } 0 < s < 2.$$

From the inequalities above and considering the finite activity jumps process, we deduce that

$$\begin{aligned}
\Delta_n^{-\beta}\,\mathrm{E}\left|\eta_{i1}^n\right| &\le K\Delta_n\left[\Delta_n^{2/3-\beta} + \Theta(\Delta_n)^{2-s}\Delta_n^{-\beta} + \Delta_n^{-1/6-\beta}\Theta(\Delta_n)^{1-s/2}\right], \\
\Delta_n^{-\beta}\,\mathrm{E}\left|\eta_{i2}^n\right| &\le K\Delta_n\left[\Theta(\Delta_n)^{2-s}\Delta_n^{-\beta} + \Delta_n^{-1/6-\beta}\Theta(\Delta_n)^{1-s/2}\right], \\
\Delta_n^{-\beta}\,\mathrm{E}\left|\eta_{i3}^n\right| &\le K\Delta_n\left[\Theta(\Delta_n)^{2-s}\Delta_n^{-\beta} + \Delta_n^{-1/6-\beta}\Theta(\Delta_n)^{1-s/2}\right].
\end{aligned}$$

By Assumption 3.1, $\lim_{\Delta_n\to 0}\Theta(\Delta_n)=0$, and $\lim_{\Delta_n\to 0}\Delta_n^{1/3}/\Theta(\Delta_n)=0$, power functions $\Theta(\Delta_n)=C\Delta_n^{\alpha}$ for any $\alpha\in(0,1/3)$ and $C\in R$ are possible choices. Now let $0<s<1$ and $\alpha\in(1/6,1/3)$, then $\beta\in(0,1/6)$ make Lemma 3.2 hold. We complete the proof of Lemma 3.2. □

Appendix E. Proof of Theorem 3.2

We define

$$\nu_i^n=\overline{Z}_i\overline{Z}_{i+k_n}1_{\{|\overline{Z}_i|\le\Theta(\Delta_n)\}}-\overline{Z}_i\overline{Z}_{i+k_n}=-\overline{Z}_i\overline{Z}_{i+k_n}1_{\{|\overline{Z}_i|>\Theta(\Delta_n)\}}.$$

By applying the Young's inequality, we get

$$\begin{aligned}\Delta_n^{-\beta}\,\mathrm{E}\left|\nu_i^n\right| &\le \frac{1}{2}\Delta_n^{-\beta}\,\mathrm{E}\left[\left(|\overline{Z}_i|^2+|\overline{Z}_{i+k_n}|^2\right)1_{\{|\overline{Z}_i|>\Theta(\Delta_n)\}}\right]\\ &=\Delta_n^{-\beta}\,\mathrm{E}\left[\left(|\overline{Y}_i|^2+2|\overline{Y}_i|\,|\overline{J}_i|+|\overline{J}_i|^2\right)1_{\{|\overline{Z}_i|>\Theta(\Delta_n)\}}\right]\\ &\le K\Delta_n\left(\Delta_n^{2/3-\beta}+\Delta_n^{5/6-\beta}+\Delta_n^{1-\beta}\right).\end{aligned}$$

Combining the proof of Lemmas 3.1 and 3.2, we complete the proof of Theorem 3.2. □