From previous-tick to pre-averaging: Spectra of equidistant transformations for unevenly spaced high-frequency data

Vitali Alexeev [1] Jun Chen [2] Katja Ignatieva [2]

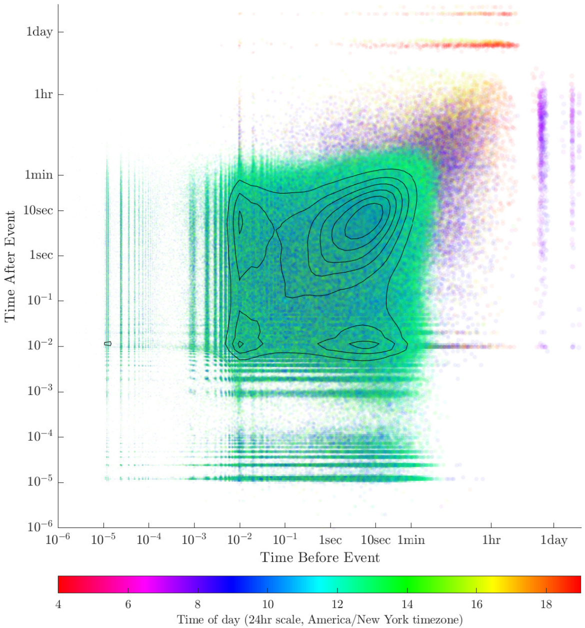
[1]UTS Business School, University of Technology Sydney, Australia
[2]School of Risk and Actuarial Studies, Business School, UNSW Australia

The 12th Bachelier World Congress
FGV EMAp, Rio de Janeiro
July 8-12, 2024

Motivation

High-frequency Data

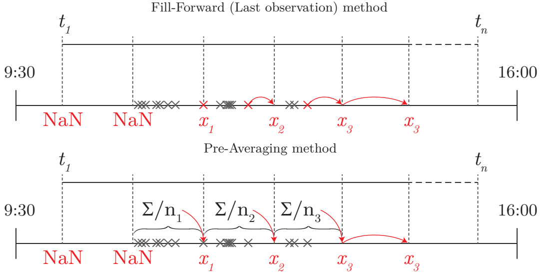
- A vast amount of data.
  - More data $\rightarrow$ More precise estimation;
  - Higher frequency $\rightarrow$ More persistent noise;
- But... (typically) Irregularly spaced
  - Tick-by-tick trade data;
  - Sentiment from textual data (news and social media);
  - Implications for univariate and, especially, multivariate analyses.


Time After Event
1day
1hr
1min
10sec
1sec
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 1sec 10sec 1min 1hr 1day
Time Before Event
4 6 8 10 12 14 16 18
Time of day (24hr scale, America/New York timezone)

Motivation

Common (calendar) schemes for equidistant data:

- *Previous tick* (also known as *last observation*)
- *Pre-average* (also recently *pre-median*, but mostly for cryptos)



Objective:

- Devise equidistant mesh to control trade-off between the number of discarded observations and the amount of (microstructure) noise.

Microstructure Noise

Let $x(t)$ be the true log price of an asset at time t, $x(t)$ satisfies

$$dx(t) = \mu(t, x(t))dt + \sigma(t, x(t))dW(t). \tag{1}$$

$$IV = \int_0^1 \sigma(t, x(t))^2 dt. \tag{2}$$

Let $y(t)$ be the observed prices, $\epsilon(t)$ be microstructure noise, it holds

$$y(t) = x(t) + \epsilon(t). \tag{3}$$

Assumption

1. $\{\epsilon(t)\}$ is stochastically independent of the price process $\{x(t)\}$;
2. $\{\epsilon(t)\}$ is white noise with $\mathrm{E}(\epsilon(t)) = 0$ and $\mathrm{Var}(\epsilon(t)) = \eta^2 < \infty$.

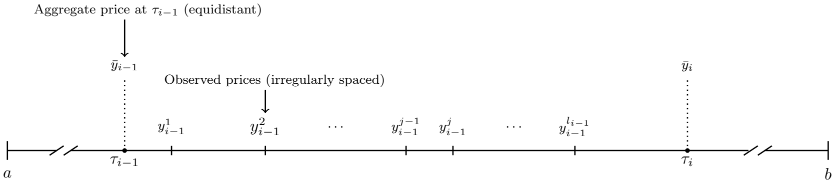
Issue: $RV \xrightarrow{p} IV, \quad n \to \infty$

Let t_i, $i = 1, \dots, n$, in $a = t_1 < t_2 < \cdots < t_n = b$ be observation times on partition $[a, b]$. An equidistant grid τ_i, $i = 1, \dots, m$, spanning the time interval $[a, b]$:

$$a = \tau_1 < \tau_2 < \cdots < \tau_m = b, \tag{4}$$

where, generally, $n > m$. The grid partitions $[a, b]$ into $m - 1$ sub-intervals.

Within a sub-interval $(\tau_{i-1}, \tau_i]$, $y^j_{\tau_{i-1}}$ is the j^{th} observation with timestamp t^j_{i-1}, and l_{i-1} is the number of obs located within the sub-interval.



Pre-weighted Sampling Scheme

Let $y_{i-1}^{1}, \ldots, y_{i-1}^{l_{i-1}}$ be the observable prices in a sub-interval $(\tau_{i-1}, \tau_i]$, and $\boldsymbol{\omega} = (\omega_{i-1}^{1}, \ldots, \omega_{i-1}^{l_{i-1}})$ be the vector of weights for these observations such that $\sum_{j=1}^{l_{i-1}} \omega_{i-1}^{j} = 1$. Then the aggregate price at τ_i:

$$\bar{y}_i = \begin{cases} \displaystyle\sum_{j=1}^{l_{i-1}} \omega_{i-1}^{j} y_{i-1}^{j}, & \text{if } \; l_{i-1} > 0, \\ \bar{y}_{i-1}, & \text{if } \; l_{i-1} = 0. \end{cases} \tag{5}$$

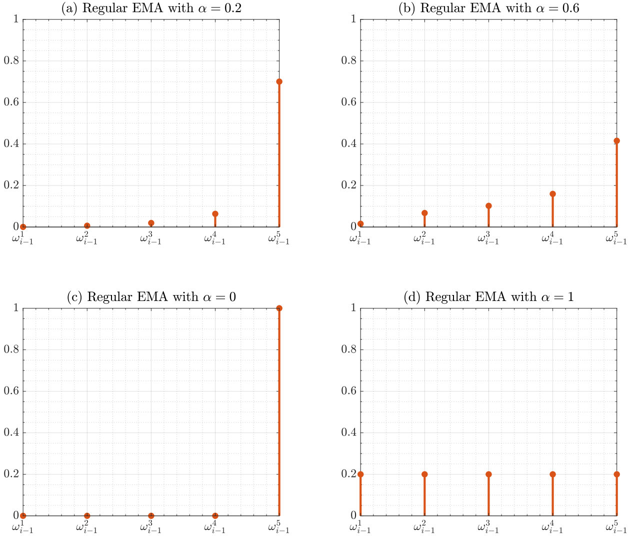
Previous tick: $y_{i-1}^{l_{i-1}}$ and Pre-average: $\dfrac{1}{l_{i-1}} \displaystyle\sum_{j=1}^{l_{i-1}} y_{i-1}^{j}$

Recursively for $j = 1, \ldots, l_{i-1}$ with initial value for $\bar{y}_i$

Regular pre-EMA: $(1 - \alpha_k) y_{i-1}^{j} + \alpha_k \bar{y}_i$

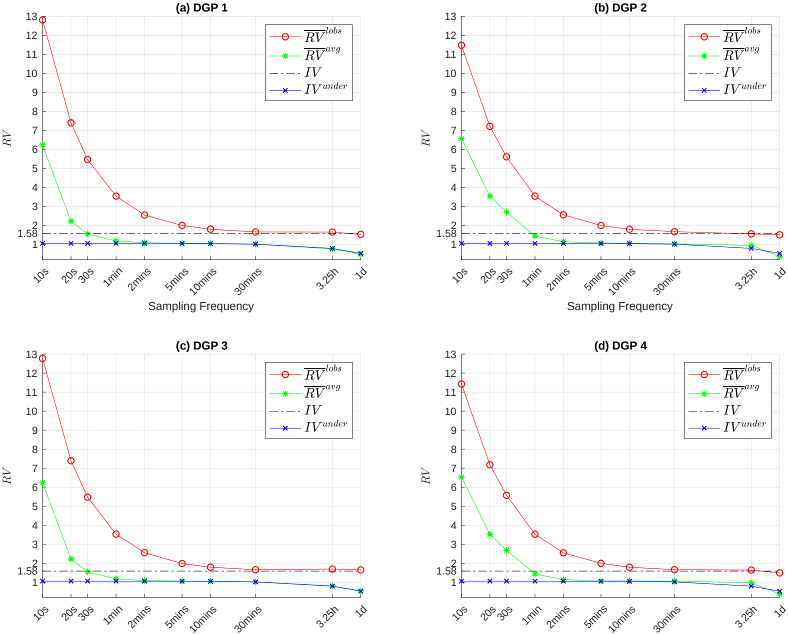
Irregular pre-EMA: $(1 - \exp(-\alpha_k \delta_j))\, y_{i-1}^{j} + \alpha_k \bar{y}_i$

Observation weights in a sub-interval for different α



Sampling Schemes: Pros and Cons

Scheme	Pros	Cons
Previous tick	·easy to implement	·discards most observations
Pre-average	·full use of all data	·equal weights to all obs ·over-emphases less recent obs ·underestimates the *IV* Details
Pre-EMA	·full use of all data ·larger weights to latest obs ·a parametric hybrid	·hyperparameter to determine Details


(a) DGP 1
(b) DGP 2
(c) DGP 3
(d) DGP 4
$\overline{RV}^{lobs}$
$\overline{RV}^{avg}$
IV
IV^{under}
RV
1.58
10s 20s 30s 1min 2mins 5mins 10mins 30mins 3.25h 1d
Sampling Frequency

Simulation Study

True price process	Sample collection method	
	Equidistant collection	Irregular collection
Brownian Motion	DGP 1	DGP 2
Heston Model	DGP 3	DGP 4

- Using Euler discretisation, simulate the continuous GBM and Heston.
- The model parameters are as in Jacod et al. (2009)
- Generate 23,401 obs each day.
- Employ two approaches to sample observations from the simulated set:
    - equidistant collection
    - irregular sample collection
- Add microstructure noise to the true prices.

Optimal smoothing parameter (varied microstructure noise)

Optimal smoothing parameter (in red) for

Regular Pre-EMA Scheme: $(1-\alpha_k)y^j_{i-1}$

Irregular Pre-EMA Scheme: $(1-\exp(-\alpha_k\delta_j))\,y^j_{i-1}$

Interpretation: $1 \rightarrow$ *Pre-averaging* and $0 \rightarrow$ *Previous Tick*

Scheme	Frequency of equidistant mesh (Δ_k)							
	10s	20s	30s	1 min	2 mins	5 mins	10 mins	30 mins
	Panel A: Large microstructure noise $\epsilon(t) \sim \mathcal{N}(0, 2*0.0005^2)$							
Regular	1	1	1	1	0.8157	0.6750	0.5795	0
Irregular	1	1	1	1	0.8395	0.7013	0.4879	0
	Panel B: Medium microstructure noise $\epsilon(t) \sim \mathcal{N}(0, 1*0.0005^2)$							
Regular	1	1	1	0.7884	0.5695	0.4655	0.2089	0
Irregular	1	1	1	0.8975	0.6537	0.6004	0.2716	0
	Panel C: Small microstructure noise $\epsilon(t) \sim \mathcal{N}(0, 0.5*0.0005^2)$							
Regular	1	1	1	0.5931	0.4248	0.3393	0	0
Irregular	1	1	1	0.6948	0.4927	0.3771	0.0461	0

Optimal smoothing parameter (varied liquidity)

Optimal smoothing parameter (in red) for

Regular Pre-EMA Scheme: $(1-\alpha_k)y^j_{i-1}$

Irregular Pre-EMA Scheme: $(1-\exp(-\alpha_k\delta_j))\,y^j_{i-1}$

Interpretation: $1 \rightarrow$ *Pre-averaging* and $0 \rightarrow$ *Previous Tick*

Scheme	Frequency of equidistant mesh (Δ_k)							
	10s	20s	30s	1 min	2 mins	5 mins	10 mins	30 mins
	Panel A: High liquidity $\overline{\lambda}_h = (60, 60, 20, 14, 6, 4, 4, 4, 20, 30, 32, 36, 50)$							
Regular	1	1	1	0.7380	0.6473	0.6385	0.5477	0.5023
Irregular	1	1	1	0.8302	0.7453	0.7272	0.5017	0
	Panel B: Intermediate $\overline{\lambda} = (40, 30, 10, 7, 3, 2, 2, 2, 10, 15, 16, 18, 25)$							
Regular	1	1	1	0.7884	0.5695	0.4655	0.2089	0
Irregular	1	1	1	0.8975	0.6537	0.6004	0.2716	0
	Panel C: Low $\overline{\lambda}_l = (20, 15, 5, 3, 1, 1, 1, 1, 5, 7, 8, 9, 12)$							
Regular	1	1	1	1	0.6114	0.3729	0.2218	0
Irregular	1	1	1	1	0.7716	0.4334	0.2282	0

Overall in-sample estimation accuracy

Using the DMW test, the number of estimators that a particular estimator can beat in terms of its estimation accuracy. MPF = marginal performance w.r.t. sampling frequencies; MPS = marginal performance w.r.t. sampling schemes.

Δ_k	$\mathrm{RV}_k^{\text{lobs}}$	$\mathrm{RV}_k^{\text{exp}}$	$\mathrm{RV}_k^{\text{iexp}}$	$\mathrm{RV}_k^{\text{ravg}}$	MPF
10s	0	4	3	1	0.0175
20s	2	10	9	6	0.0591
30s	5	12	11	7	0.0766
1min	8	28	28	18	0.1794
2min	13	28	28	25	0.2057
5min	19	24	24	24	0.1991
10min	19	19	19	18	0.1641
30min	12	11	11	11	0.0985
MPS	0.1707	0.2976	0.2910	0.2407	1.0000

Overall out-of-sample estimation accuracy

Using the DMW test, the number of estimators that a particular estimator can beat in terms of its estimation accuracy. MPF = marginal performance w.r.t. sampling frequencies; MPS = marginal performance w.r.t. sampling schemes.

Δ_k	$\mathrm{RV}_k^{\mathsf{lobs}}$	$\mathrm{RV}_k^{\mathsf{exp}}$	$\mathrm{RV}_k^{\mathsf{iexp}}$	$\mathrm{RV}_k^{\mathsf{ravg}}$	MPF
10s	0	4	3	1	0.0175
20s	2	10	9	6	0.0590
30s	5	12	11	7	0.0764
1min	8	28	28	18	0.1790
2min	13	28	28	24	0.2031
5min	19	24	24	24	0.1987
10min	19	19	19	18	0.1638
30min	13	11	12	11	0.1026
MPS	0.1725	0.2969	0.29265	0.2380	1.0000

Data Summary and Optimal smoothing parameter

- Tick-by-tick trades for NASDAQ:GOOG.
- January 2005 to December 2014.
- 45,951,783 observations across 2,513 days (after data cleaning).
- i.e., 18,285 obs per day.
- Two sub-samples, allowing performance comparison by means of in- and out-of-sample tests.

Optimal smoothing parameter (in red) for
Regular Pre-EMA Scheme: $(1 - \alpha_k) y^j_{i-1}$
Irregular Pre-EMA Scheme: $(1 - \exp(-\alpha_k \delta_j))\, y^j_{i-1}$
Interpretation: 1 → *Pre-averaging* and 0 → *Previous Tick*

Scheme	Frequency of equidistant mesh (Δ_k)							
	10s	20s	30s	1min	2min	5min	10min	30min
Regular	0.7411	0.4847	0.3740	0.1552	0	0	0	0
Irregular	0.7015	0.4239	0.3124	0.0776	0	0	0	0

Overall in-sample estimation accuracy

Using the DMW test, the number of estimators that a particular estimator can beat in terms of its estimation accuracy. MPF = marginal performance w.r.t. sampling frequencies; MPS = marginal performance w.r.t. sampling schemes.

Δ_k	$\mathrm{RV}_k^{\mathrm{lobs}}$	$\mathrm{RV}_k^{\mathrm{exp}}$	$\mathrm{RV}_k^{\mathrm{iexp}}$	$\mathrm{RV}_k^{\mathrm{ravg}}$	MPF
10s	4	16	16	9	0.1119
20s	11	28	25	21	0.2114
30s	16	24	24	20	0.2090
1min	20	20	20	15	0.1866
2min	16	15	15	13	0.1468
5min	9	8	8	8	0.0821
10min	5	4	4	4	0.0423
30min	1	0	0	3	0.0100
MPS	0.2040	0.2861	0.2786	0.2313	1.0000

Overall out-of-sample estimation accuracy

Using the DMW test, the number of estimators that a particular estimator can beat in terms of its estimation accuracy. MPF = marginal performance w.r.t. sampling frequencies; MPS = marginal performance w.r.t. sampling schemes.

Δ_k	$\mathrm{RV}_k^{\mathrm{lobs}}$	$\mathrm{RV}_k^{\mathrm{exp}}$	$\mathrm{RV}_k^{\mathrm{iexp}}$	$\mathrm{RV}_k^{\mathrm{ravg}}$	MPF
10s	8	17	17	14	0.1425
20s	13	26	22	28	0.2265
30s	17	20	20	21	0.1985
1min	17	17	17	16	0.1705
2min	13	12	12	12	0.1247
5min	9	8	8	8	0.0840
10min	5	4	4	4	0.0434
30min	1	0	0	3	0.0102
MPS	0.2112	0.2646	0.2545	0.2697	1.0000

Summary

- Connect the *previous tick sampling scheme* and the *pre-averaging sampling scheme* by the *pre-EMA sampling scheme*;
- Apply *pre-EMA sampling scheme* to irregular high frequency data;
- Develop the method to find the optimal α in the *regular pre-EMA sampling scheme* and the method can also be applied to the *irregular pre-EMA sampling scheme*;
- Conduct comparison between RVs at various sampling frequencies with different sampling schemes.

Thanks!

Pre-averaging Sampling Scheme

Back to Main

Assumption 3

There are L observations in each sub-interval $(\tau_{i-1}, \tau_i]$ for $i = 1, \cdots, m$.

Lemma

Under stated assumptions, it holds

$$E(RV^{avg}) = \frac{1 + 2L^2}{3L^2} IV + 2\frac{m}{L}\eta^2 \tag{6}$$

where, RV^{avg} is the RV based on the pre-averaging sampling scheme. Thus

$$\lim_{L \to \infty} E(RV^{avg}) = \frac{2}{3} IV \tag{7}$$

Optimal α for Regular Pre-EMA Sampling Scheme

Back to Main

Let θ_n be the true value of IV, $\tilde{\theta}_n$ be the noisy but unbiased estimator of θ_n, we assume

1. $\tilde{\theta}_n = \theta_n + \nu_n$, with $\mathbb{E}[\nu_n | \mathcal{F}_{n-1}, \theta_n] = 0$;
2. $\theta_n = \theta_{n-1} + \vartheta_n$, with $\mathbb{E}[\vartheta_n | \mathcal{F}_{n-1}] = 0$;
3. $\Upsilon_n = \sum_{j=1}^{J} \lambda_j \tilde{\theta}_{n+j}$, where $1 \leq J < \infty$, $\lambda_j \geq 0 \quad \forall j$ and $\sum_{j=1}^{J} \lambda_j = 1$.

Optimal α minimizes the objective function

$$\alpha^* \equiv \arg\min_{\alpha} \mathbb{E}[Q(RV_{\alpha}^{exp}(n), \theta_n)] \tag{8}$$

or, equivalently:

$$\tilde{\alpha}^* \equiv \arg\min_{\tilde{\alpha}} \mathbb{E}[Q(RV_{\alpha}^{exp}(n), \tilde{\theta}_n)]. \tag{9}$$

Difference in Size of Microstructure Noise

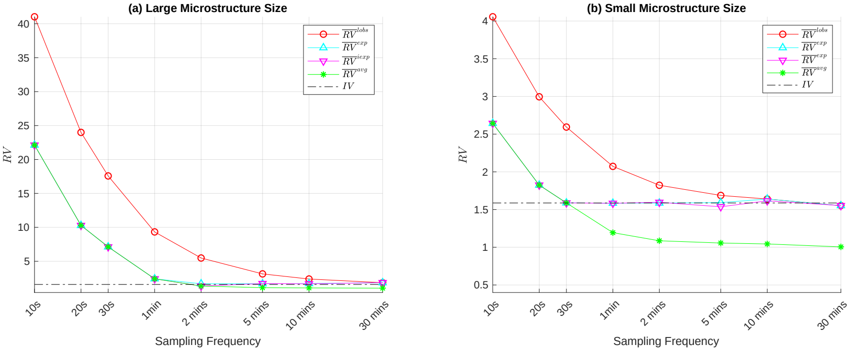
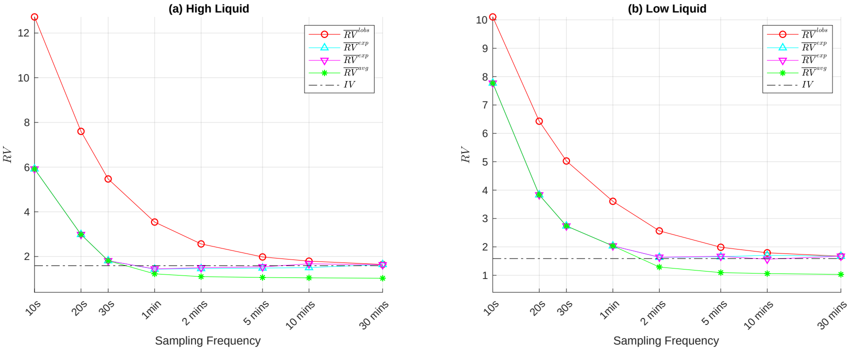


Figure: $\epsilon(t) \sim \mathcal{N}(0, 2 * 0.0005^2)$ v.s. $\epsilon(t) \sim \mathcal{N}(0, 0.5 * 0.0005^2)$

Difference in Liquidity



Diebold-Mariano-West tests results (5%)

"✓" = RV from the row has greater estimation accuracy than RV from the column.
"✗" = reject the null hypothesis in favour of the RV from the corresponding column.
"-" = both estimators are comparable in terms of estimation accuracy.

(A) RV_k^{lobs} estimator.

$RV_{(\dots)}^{lobs}$	10s	20s	30s	1min	2min	5min	10min	30min	1d
10s	-	✗	✗	✗	✗	✗	✗	✗	✓
20s	✓	-	✗	✗	✗	✗	✗	✗	✓
30s	✓	✓	-	✗	✗	✗	✗	✗	✓
1min	✓	✓	-	-	✗	✗	✗	✗	✓
2min	✓	✓	✓	✓	-	✗	✗	-	✓
5min	✓	✓	✓	✓	✓	-	-	✓	✓
10min	✓	✓	✓	✓	✓	-	-	✓	✓
30min	✓	✓	✓	✓	-	✗	✗	-	✓
1d	✗	✗	✗	✗	✗	✗	✗	✗	-

(B) RV_k^{exp} estimator.

$RV_{(\dots)}^{exp}$	10s	20s	30s	1min	2min	5min	10min	30min
10s	-	✗	✗	✗	✗	✗	✗	✗
20s	✓	-	✗	✗	✗	✗	✗	✗
30s	✓	✓	-	✗	✗	✗	✗	-
1min	✓	✓	✓	-	-	✓	✓	✓
2min	✓	✓	✓	-	-	✓	✓	✓
5min	✓	✓	✓	✗	✗	-	✓	✓
10min	✓	✓	✓	✗	✗	✗	-	✓
30min	✓	✓	-	✗	✗	✗	✗	-

(C) RV_k^{iexp} estimator.

$RV_{(\dots)}^{iexp}$	10s	20s	30s	1min	2min	5min	10min	30min
10s	-	✗	✗	✗	✗	✗	✗	✗
20s	✓	-	✗	✗	✗	✗	✗	✗
30s	✓	✓	-	✗	✗	✗	✗	-
1min	✓	✓	✓	-	-	✓	✓	✓
2min	✓	✓	✓	-	-	✓	✓	✓
5min	✓	✓	✓	✗	✗	-	✓	✓
10min	✓	✓	✓	✗	✗	✗	-	✓
30min	✓	✓	-	✗	✗	✗	✗	-

(D) RV_k^{ravg} estimator.

$RV_{(\dots)}^{ravg}$	10s	20s	30s	1min	2min	5min	10min	30min
10s	-	✗	✗	✗	✗	✗	✗	✗
20s	✓	-	✗	✗	✗	✗	✗	✗
30s	✓	✓	-	✗	✗	✗	✗	✗
1min	✓	✓	✓	-	✗	✗	-	✓
2min	✓	✓	✓	✓	-	✓	✓	✓
5min	✓	✓	✓	✓	✗	-	✓	✓
10min	✓	✓	✓	-	✗	✗	-	✓
30min	✓	✓	✓	✗	✗	✗	✗	-

Diebold-Mariano-West tests results (5%) - GOOG.OQ

"✓" = RV from the row has greater estimation accuracy than RV from the column.
"✗" = reject the null hypothesis in favour of the RV from the corresponding column.
"-" = both estimators are comparable in terms of estimation accuracy.

(E) $\mathrm{RV}_k^{\mathrm{lobs}}$ estimator.

$\mathrm{RV}_{(\dots)}^{\mathrm{lobs}}$	10s	20s	30s	1min	2min	5min	10min	30min	1d
10s	-	✗	✗	✗	✗	-	-	✓	-
20s	✓	-	✗	✗	-	-	✓	✓	-
30s	✓	✓	-	-	-	✓	✓	✓	✓
1min	✓	✓	-	-	✓	✓	✓	✓	✓
2min	✓	-	-	✗	-	✓	✓	✓	✓
5min	-	-	✗	✗	✗	-	✓	✓	✓
10min	-	✗	✗	✗	✗	✗	-	✓	✗
30min	✗	✗	✗	✗	✗	✗	✗	-	✗
1d	-	-	✗	✗	✗	✗	✓	✓	-

(F) $\mathrm{RV}_k^{\mathrm{exp}}$ estimator.

$\mathrm{RV}_{(\dots)}^{\mathrm{exp}}$	10s	20s	30s	1min	2min	5min	10min	30min
10s	-	✗	-	-	-	✓	✓	✓
20s	✓	-	-	✓	✓	✓	✓	✓
30s	-	-	-	✓	✓	✓	✓	✓
1min	-	✗	✗	-	✓	✓	✓	✓
2min	-	✗	✗	✗	-	✓	✓	✓
5min	✗	✗	✗	✗	✗	-	✓	✓
10min	✗	✗	✗	✗	✗	✗	-	✓
30min	✗	✗	✗	✗	✗	✗	✗	-

(G) $\mathrm{RV}_k^{\mathrm{iexp}}$ estimator.

$\mathrm{RV}_{(\dots)}^{\mathrm{iexp}}$	10s	20s	30s	1min	2min	5min	10min	30min
10s	-	✗	-	-	-	✓	✓	✓
20s	✓	-	-	-	✓	✓	✓	✓
30s	-	-	-	✓	✓	✓	✓	✓
1min	-	-	✗	-	✓	✓	✓	✓
2min	-	✗	✗	✗	-	✓	✓	✓
5min	✗	✗	✗	✗	✗	-	✓	✓
10min	✗	✗	✗	✗	✗	✗	-	✓
30min	✗	✗	✗	✗	✗	✗	✗	-

(H) $\mathrm{RV}_k^{\mathrm{ravg}}$ estimator.

$\mathrm{RV}_{(\dots)}^{\mathrm{ravg}}$	10s	20s	30s	1min	2min	5min	10min	30min
10s	-	✗	✗	✗	-	-	✓	✓
20s	✓	-	-	✓	✓	✓	✓	✓
30s	✓	-	-	✓	✓	✓	✓	✓
1min	✓	✗	✗	-	✓	✓	✓	✓
2min	-	✗	✗	✗	-	✓	✓	✓
5min	-	✗	✗	✗	✗	-	✓	✓
10min	✗	✗	✗	✗	✗	✗	-	✓
30min	✗	✗	✗	✗	✗	✗	✗	-