Statistical error bounds for weighted mean and median, with application to robust aggregation of cryptocurrency data

M. Allouche M. Echenim E. Gobet A-C. Maurice

Blockchain@X-OMI Workshop on Blockchain and Decentralized Finance

2023, September 21st

https://hal.science/hal-04017151v1/document

Introduction

Context.

- Crypto assets are traded in a **multi-fragmented** market
- Kaiko's DB: 2000 **assets**, 6000 **pairs**, 100 **exchanges**

Introduction

Context.

- Crypto assets are traded in a **multi-fragmented** market
- Kaiko's DB: 2000 **assets**, 6000 **pairs**, 100 **exchanges**

State of the art. Given price and volume observations $(P_i, V_i)_{i=1}^n$

- Average Price (AP) [Vinter, 2021]

$$\frac{1}{n}\sum_{i=1}^{n} P_i.$$

- Volume-Weighted Average Price (VWAP) [Nasdaq, 2022, FTSE Digital Asset Research, 2022]

$$\widehat{\texttt{VWAP}}_n := \frac{\sum_{i=1}^n V_i P_i}{\sum_{i=1}^n V_i}.$$

- Volume-Weighted Median Price (VWM)
  [Federal Reserve Bank of New York, 2015, CF Benchmarks, 2022, Gallagher, 2018]

$$\widehat{\texttt{VWM}}_n := \inf\left\{p : \frac{\sum_{i=1}^n V_i \mathbb{1}_{P_i \le p}}{\sum_{i=1}^n V_i} \ge \frac{1}{2}\right\}.$$

Problem Statement

Challenge. Large **discrepancies** in price, volume, liquidity among the exchanges make the task of "**market consensus**" price calculation difficult

Objective.

- Study the statistical distribution of the crypto data (price/volume)
- Control the statistical fluctuations of the methods for both **small** and **large** datasets
- Propose a new challenger : **Robust Weighted Median** (**RWM**)

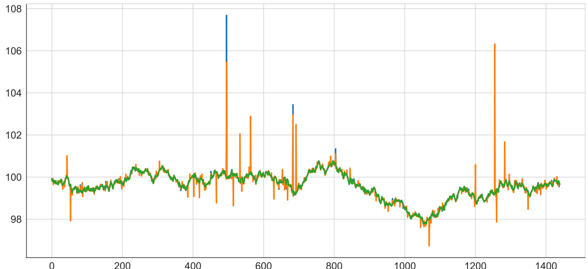
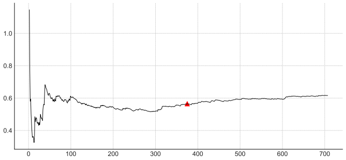


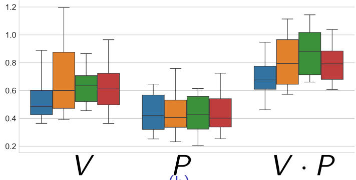
Figure: Aggregated volume weighted price estimation on simulated data surrounding an efficient price generated every minute during one day using $\widehat{\texttt{VWM}}_n$ (blue), $\widehat{\texttt{VWAP}}_n$ (orange) and $\widehat{\texttt{RWM}}_n$ (green).

Data analysis - Price and Volume are heavy-tailed

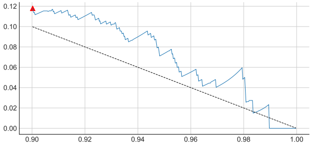
- Check the tail behavior of (P, V) : 67 **pairs**, 4 **days**, > 80**M trades**
- Heavy-tailed distributions are characterized by a **survival function** which decays at a rate $x^{-1/\gamma}$ as $x \to \infty$, where $\gamma > 0$ is called the **tail index**



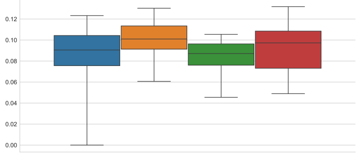
(a)



(b)



(c)



(d)

Figure: (a): Hill plot [Hill, 1975] regarding the price returns btc/usdc on 2022-05-05. (b): Box plot of the estimated tail-index on the volumes (left), on the price returns (middle) and on the product of the two (right) for all considered pairs on 2022-05-05 (blue), 2022-06-12 (orange), 2022-06-28 (green) and 2022-12-09 (red). (c): Upper tail dependence estimator between price returns and volumes (blue), and the independence case (black dashed line) for btc/usdc on 2022-05-05. (d): Box plot of the estimated upper tail dependence for all considered pairs and periods

Statistical analysis

Consider two positive r.v. P, W and the independent observations $(P_i, W_i)_{i=1}^n$ drawn from (P, W). Introduce the model and data based definitions:

c.d.f.

$$F_{\mathtt{W}}(x) := \frac{\mathbb{E}\left[W \cdot \mathbb{1}_{P \leq x}\right]}{\mathbb{E}\left[W\right]}, \qquad \widehat{F_{\mathtt{W,n}}}(x) := \frac{\sum_{i=1}^n W_i \mathbb{1}_{P_i \leq x}}{\sum_{i=1}^n W_i}$$

Weighted Average Price.

$$\mathtt{WAP} := \frac{\mathbb{E}\left[W \cdot P\right]}{\mathbb{E}\left[W\right]}, \qquad \widehat{\mathtt{WAP}}_n := \frac{\sum_{i=1}^n W_i P_i}{\sum_{i=1}^n W_i}$$

Weighted Quantile.

$$q_{\mathtt{W}}(\alpha) := \inf\left\{p : \frac{\mathbb{E}\left[W \mathbb{1}_{P \leq p}\right]}{\mathbb{E}\left[W\right]} \geq \alpha\right\}, \qquad \widehat{q_{\mathtt{W},n}}(\alpha) := \inf\left\{p : \frac{\sum_{i=1}^n W_i \mathbb{1}_{P_i \leq p}}{\sum_{i=1}^n W_i} \geq \alpha\right\}$$

Weighted Median.

$$\mathtt{WM} := q_{\mathtt{W}}\left(\frac{1}{2}\right), \qquad \widehat{\mathtt{WM}}_n := \widehat{q_{\mathtt{W},n}}\left(\frac{1}{2}\right)$$

Assumptions

(H0): The c.d.f. $F_{\mathtt{W}}$ in has a density $f_{\mathtt{W}}$:

$$F_{\mathtt{W}}(x) = \int_0^x f_{\mathtt{W}}(x')\,\mathrm{d}x'.$$

$(\mathbf{H}_{\mathbf{X}}^{\kappa})$: X has a finite moment of order $\kappa > 2$: $\mathbb{E}\left[X^{\kappa}\right] < +\infty$.

$(\mathbf{H}_{\mathbf{X}}^{\boldsymbol{\Gamma}})$: X has a sub-gamma distribution, i.e. $\mathbb{E}\left[e^{cX}\right] < +\infty$ for some $c > 0$.

$(\mathbf{H}_{\mathbf{X}}^{\mathbf{G}})$: X has a sub-Gaussian distribution, i.e. $\mathbb{E}\left[e^{cX^2}\right] < +\infty$ for some $c > 0$.

Asymptotic fluctuations

Theorem

Assume **(H0)** *with a continuous and non-vanishing density* $f_{\mathtt{W}}$ *at* $q_{\mathtt{W}}(\alpha)$, *and assume* $(\mathbf{H}_{\mathbf{X}}^{\kappa})$ *for* $X = W$ *and some* $\kappa > 2$. *Then*

$$\sqrt{n}\left(\widehat{q_{\mathtt{W},n}}(\alpha) - q_{\mathtt{W}}(\alpha)\right) \Rightarrow \mathcal{N}\left(0, \frac{\mathbb{E}\left[W^2\left(\alpha - \mathbb{1}_{P \leq q(\alpha)}\right)^2\right]}{\left(\mathbb{E}\left[W\right] f_{\mathtt{W}}\left(q_{\mathtt{W}}(\alpha)\right)\right)^2}\right). \tag{1}$$

Assume $(\mathbf{H}_{\mathbf{X}}^{\kappa})$ *for* $X = W$ *and* $X = W \cdot P$ *and for some* $\kappa > 2$, *then*

$$\sqrt{n}\left(\widehat{\mathtt{WAP}}_n - \mathtt{WAP}\right) \Rightarrow \mathcal{N}\left(0, \frac{\mathbb{E}\left[W^2(\mathtt{WAP} - P)^2\right]}{\left(\mathbb{E}\left[W\right]\right)^2}\right). \tag{2}$$

Non-asymptotic fluctuations

$$\mathcal{Q}_n^{>}(\alpha, x) := \mathbb{P}\left(\widehat{q_{\mathtt{W},n}}(\alpha) - q_{\mathtt{W}}(\alpha) > \frac{x}{\sqrt{n}}\right), \quad \mathcal{Q}_n^{\leq}(\alpha, x) := \mathbb{P}\left(\widehat{q_{\mathtt{W},n}}(\alpha) - q_{\mathtt{W}}(\alpha) \leq -\frac{x}{\sqrt{n}}\right),$$

$$\mathcal{W}_n^{\geq}(x) := \mathbb{P}\left(\widehat{\mathtt{WAP}}_n - \mathtt{WAP} \geq \frac{x}{\sqrt{n}}\right), \qquad \mathcal{W}_n^{\leq}(x) := \mathbb{P}\left(\widehat{\mathtt{WAP}}_n - \mathtt{WAP} \leq -\frac{x}{\sqrt{n}}\right).$$

Theorem (simplified)

	$\max(\mathcal{Q}_n^{>}(\alpha, x), \mathcal{Q}_n^{\leq}(\alpha, x))$	$\max(\mathcal{W}_n^{\geq}(x), \mathcal{W}_n^{\leq}(x))$
Heavy-Tail assumptions *($\mathbf{H}_{\mathbf{X}}^{\kappa}$) for $X = W$ and $X = W \cdot P$, for $\kappa > 2$*	$\frac{c}{n^{\frac{\kappa}{2}-1} x^{\kappa}} + \exp\left(-cx^2\right)$	$\frac{c\left(1 + \frac{x}{\sqrt{n}}\right)^{\kappa}}{n^{\frac{\kappa}{2}-1} x^{\kappa}} + \exp\left(-\frac{cx^2}{1 + c\frac{x^2}{n}}\right)$
Sub-gamma assumptions *(**H0**) and ($\mathbf{H}_{\mathbf{X}}^{\Gamma}$) for $X = W$ and $X = W \cdot P$*	$\exp\left(-\frac{cx^2}{1 + c\frac{x}{\sqrt{n}}}\right)$	$\exp\left(-\frac{cx^2}{\left(1 + c\frac{x}{\sqrt{n}}\right)^3}\right)$
Sub-Gaussian assumptions *(**H0**) and ($\mathbf{H}_{\mathbf{X}}^{\mathbf{G}}$) for $X = W$ and $X = W \cdot P$*	$\exp\left(-cx^2\right)$	$\exp\left(-\frac{cx^2}{1 + c\frac{x^2}{n}}\right)$

Non-asymptotic fluctuations

$$\mathcal{Q}_n^{>}(\alpha,x) := \mathbb{P}\left(\widehat{q_{\mathtt{W},n}}(\alpha) - q_{\mathtt{W}}(\alpha) > \frac{x}{\sqrt{n}}\right), \quad \mathcal{Q}_n^{\leq}(\alpha,x) := \mathbb{P}\left(\widehat{q_{\mathtt{W},n}}(\alpha) - q_{\mathtt{W}}(\alpha) \leq -\frac{x}{\sqrt{n}}\right),$$

$$\mathcal{W}_n^{\geq}(x) := \mathbb{P}\left(\widehat{\mathtt{WAP}}_n - \mathtt{WAP} \geq \frac{x}{\sqrt{n}}\right), \qquad \mathcal{W}_n^{\leq}(x) := \mathbb{P}\left(\widehat{\mathtt{WAP}}_n - \mathtt{WAP} \leq -\frac{x}{\sqrt{n}}\right).$$

Theorem (simplified)

	$\max(\mathcal{Q}_n^{>}(\alpha,x), \mathcal{Q}_n^{\leq}(\alpha,x))$	$\max(\mathcal{W}_n^{\geq}(x), \mathcal{W}_n^{\leq}(x))$
Heavy-Tail assumptions $(\mathbf{H}_{\mathbf{X}}^{\kappa})$ *for* $X = W$ *and* $X = W \cdot P$, *for* $\kappa > 2$	$\frac{c}{n^{\frac{\kappa}{2}-1} x^{\kappa}} + \exp\left(-cx^2\right)$	$\frac{c\left(1+\frac{x}{\sqrt{n}}\right)^{\kappa}}{n^{\frac{\kappa}{2}-1} x^{\kappa}} + \exp\left(-\frac{cx^2}{1+c\frac{x^2}{n}}\right)$
Sub-gamma assumptions $(\mathbf{H0})$ *and* $(\mathbf{H}_{\mathbf{X}}^{\Gamma})$ *for* $X = W$ *and* $X = W \cdot P$	$\exp\left(-\frac{cx^2}{1+c\frac{x}{\sqrt{n}}}\right)$	$\exp\left(-\frac{cx^2}{\left(1+c\frac{x}{\sqrt{n}}\right)^3}\right)$
Sub-Gaussian assumptions $(\mathbf{H0})$ *and* $(\mathbf{H}_{\mathbf{X}}^{\mathbf{G}})$ *for* $X = W$ *and* $X = W \cdot P$	$\exp\left(-cx^2\right)$	$\exp\left(-\frac{cx^2}{1+c\frac{x^2}{n}}\right)$

▷ Define $\mathtt{RWM}$ by setting $W = \log(1 + V/q_{0.5}(V))$ which is **sub-gamma** distributed

Experiments - simulated data

- Simulate an efficient price series $(S_t)_{t=1}^{T}$ where $S_t \sim GBM(\sigma)$
- Simulate the dependent variables (P, V) with an Archimedean copula
- At each time step t, simulate J noisy prices s.t. $\forall j \in \{1, \ldots, J\}$, $\widetilde{S}_t^j = S_t(1 + r_t^j)$ where $r_t^j \sim \mathtt{Mixt}(\omega)$.

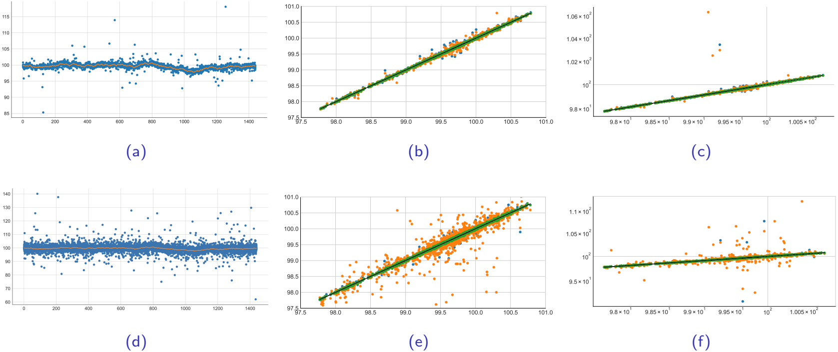


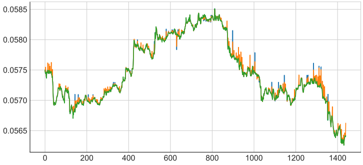
Figure: (a)-(d): Efficient price time-series $(S_t)_{t=1}^{1440}$ (orange line) simulated with $\sigma = 0.5$ ($\sigma^{\mathrm{eff}} = 0.51$), and its associated noisy prices $(\widetilde{S}_t^j, j \in \{1, \ldots, 100\})_{t=1}^{1440}$ (blue dots) with $\omega = 0.99$ (a) and $\omega = 0.7$ (d). (b)-(e): Scatter plots of the associated estimated prices $(\widehat{S}_t)_{t=1}^{100}$ ($\widehat{\mathtt{VWM}}_n$: blue, $\widehat{\mathtt{VWAP}}_n$: orange, $\widehat{\mathtt{RWM}}_n$: green) with respect to the reference price. Black dashed regression line $x \mapsto y = x$. (c)-(f): Scatter plots with log-scale axes.

Results

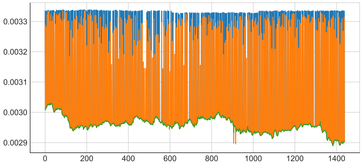
	RMSE$^{\texttt{price}}$			RMSE$^{\texttt{RV}}$		
ω	$\widehat{\texttt{VWM}}_n$	$\widehat{\texttt{VWAP}}_n$	$\widehat{\texttt{RWM}}_n$	$\widehat{\texttt{VWM}}_n$	$\widehat{\texttt{VWAP}}_n$	$\widehat{\texttt{RWM}}_n$
0.99	3.0	3.7	**1.3**	28.1	4.8	**0.02**
0.95	3.1	32.2	**1.4**	52.5	8.5	**0.02**
0.90	3.2	36.4	**1.5**	63.0	10.4	**0.02**
0.80	3.5	43.8	**1.6**	74.7	13.0	**0.03**
0.70	37.9	53.0	**1.8**	120.6	18.1	**0.03**
0.60	38.2	62.9	**2.1**	141.7	21.1	**0.04**
0.50	97.8	115.5	**2.5**	168.3	23.7	**0.06**
0.40	116.3	127.7	**3.0**	187.1	26.1	**0.09**
0.30	116.5	131.4	**3.9**	201.31	28.25	**0.14**
0.20	122.6	134.0	**5.5**	222.5	30.9	**0.26**
0.10	123.0	136.0	**8.2**	226.1	31.9	**0.51**

Table: Comparison between $\widehat{\texttt{VWM}}_n$, $\widehat{\texttt{VWAP}}_n$ and $\widehat{\texttt{RWM}}_n$ results for different simulation scenarios with $\omega \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$ using two performance criteria.

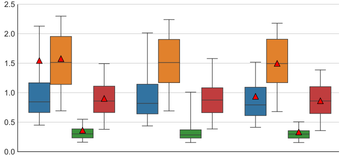
Experiments - real data



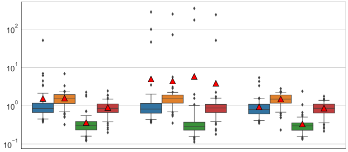
(a)



(b)



(c)



(d)

Figure: (a)-(b): Comparison between $\widehat{\texttt{VWM}}_n$ (blue), $\widehat{\texttt{VWAP}}_n$ (orange) and $\widehat{\texttt{RWM}}_n$ (green) on real data aggregation per minute of the pair eth/btc (a) and zec/btc (b) on 2022-06-28. (c): Box plot of the annualized RV (over all pairs) from $\widehat{\texttt{VWM}}_n$ (left), $\widehat{\texttt{VWAP}}_n$ (middle) and $\widehat{\texttt{RWM}}_n$ (right) on 2022-06-28 (blue), on 2022-06-12 (orange), on 2022-09-12 (green) and on 2022-05-05 (red). The mean is emphasized by a red triangle and outliers are discarded. d)Box plot of the annualized RV taking into account outliers with the y-axis in log-scale.

RWM outperforms VWAP and VWM - SVB crisis

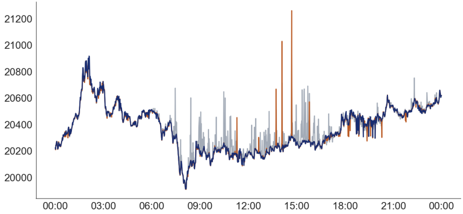


Figure: **BTC-USD** on 2023/03/11

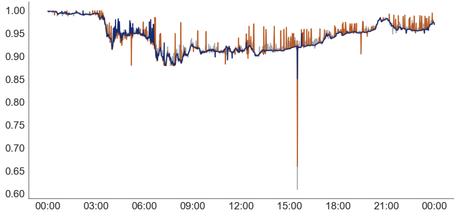


Figure: **USDC-USD** on 2023/03/11

Conclusion

- Provide theoretical results on both **asymptotic** and **non-asymptotic bounds** of the estimators.
- Show that VWAP and VWM suffer from **instability** and **lack of robustness** when applied to crypto data that are **heavy-tailed**
- **Outperform** other competitors in both simulated and real data

References I

CF Benchmarks (2022).
CME CF cryptocurrency reference rates, version 15.1.
https://docs.cfbenchmarks.com/CME%20CF%20Reference%20Rates%20Methodology.pdf.

Federal Reserve Bank of New York (2015).
Technical note concerning the methodology for calculating the effective federal funds rate.
https://www.newyorkfed.org/medialibrary/media/markets/EFFR-technical-note-070815.pdf.

FTSE Digital Asset Research (2022).
Guide to the calculation of the FTSE DAR Digital Asset Prices and FTSE DAR Reference Prices.
https://research.ftserussell.com/products/downloads/Guide_to_the_Calculation_of_FTSE_DAR_Digital_Asset_Prices_and_Reference_Prices_Fixes.pdf.

References II

Gallagher, C. (2018).
CFIX methodology, Bloomberg cryptocurrency solutions.
https://data.bloomberglp.com/professional/sites/10/CFIX-Methodology.pdf.

Hill, B. M. (1975).
A simple general approach to inference about the tail of a distribution.
Ann. Statist., 3(5):1163–1174.

Nasdaq (2022).
Nasdaq Crypto Index: index methodology.
https://indexes.nasdaqomx.com/docs/methodology_NCI.pdf.

Vinter (2021).
Crypto reference rates for single assets, version 2.1.
https://methodology.vinter.co/vinter/reference-rates.