Robust aggregation of crypto data

Michaël ALLOUCHE, Mnacho ECHENIM,
Emmanuel GOBET, Anne-Claire MAURICE

ICCF24 - CWI, Amsterdam
April 5th, 2024

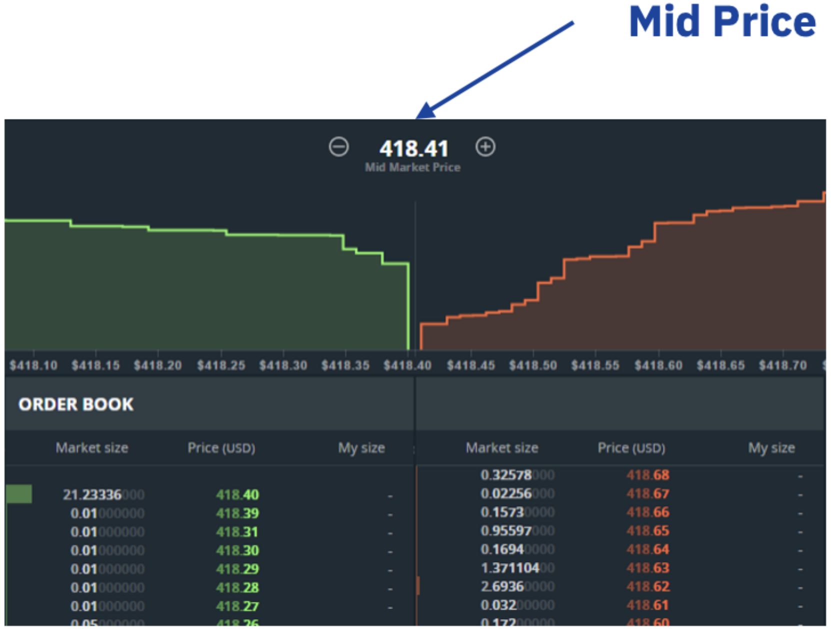
Preprint available on HAL server:
https://hal.science/hal-04017151
Tentatively accepted for publication in *Mathematical Finance*

1 Pb statement: Price Discovery

▷ Limit Order Book in Traditional Finance

Mid Price

Makers
Provide liquidity by placing sell orders



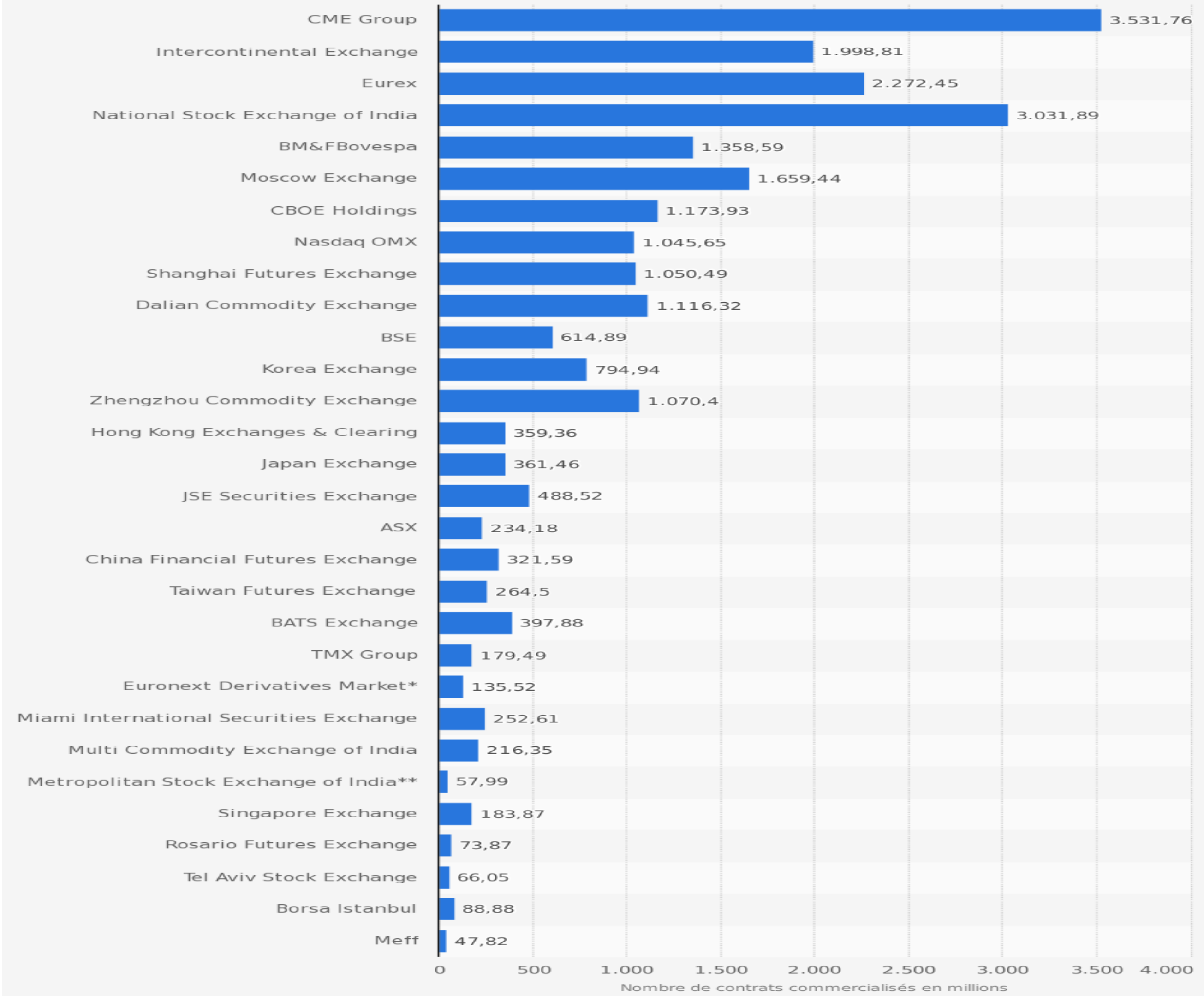
Takers
Take liquidity by placing buy orders

A trade occurs if the buy / sell orders match up

▷ One asset = one exchange = one price at a given time

✓ Traditional Finance: a few big exchanges

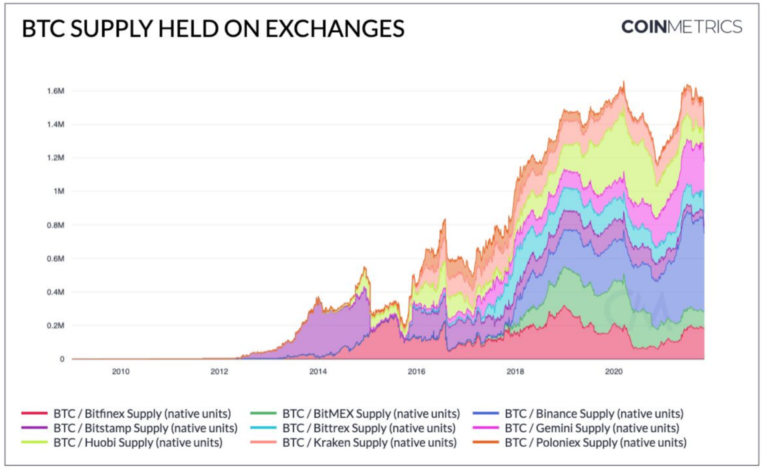
Main exchanges for derivatives in the work in 2015, in terms of number of contracts (in millions) - Source: Futures Industry Association



✓ Crypto Finance: **fragmented market**

- Centralized Exchanges (off-chain)

Hundreds exist around the world



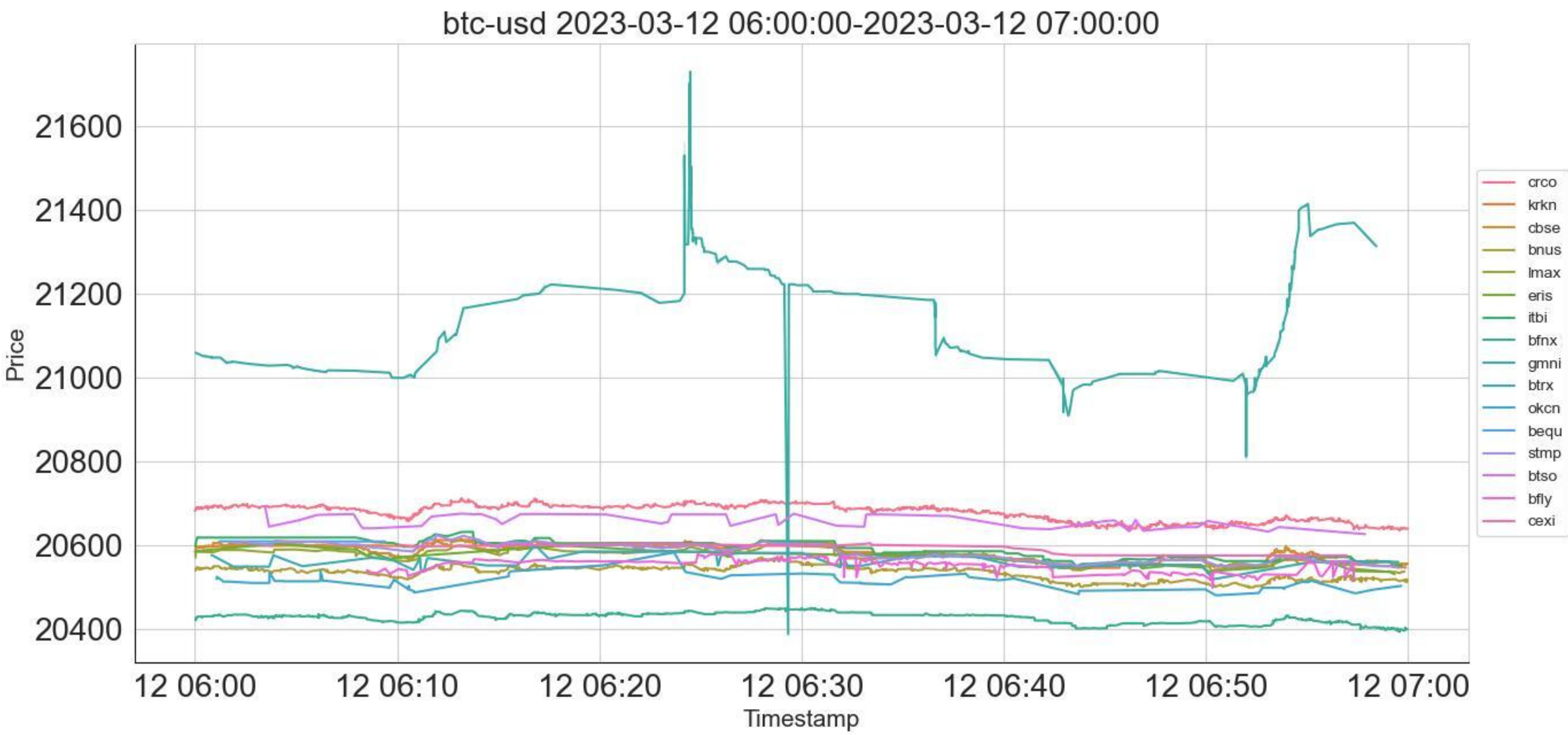
- Decentralized Exchanges (on-chain):

# Trades	Exchange	# Pools	%	# Mints	# Burns
960924	curv	344	0,16%	566152	421482
1398792	blc2	1275	0,59%	93864	44279
471382	crv2	18	0,01%	18261	11276
114814668	usp2	195403	90,03%	3924111	1740615
11976425	sush	3684	1,70%	531695	308206
32452350	usp3	13037	6,01%	1108823	1242071
2787827	blcr	3291	1,52%	305293	200417

Number of trades and liquidity pools, by decentralized exchange, on Ethereum

- About 1700 cryptocurrencies traded on a daily basis

✓ **What is the "market consensus" price for a digital asset?**



Statistical problem: estimate the market consensus price

- ✓ Data for a given pair of assets:
  - list of exchanges
  - for each, volume and price of trades at different timestamps
- ✓ In such a fragmented market, how to compute a "market price consensus" at a given date? with high frequency update?
- ✓ Key desirable properties of a price aggregator [PK16]:
  - Relevance,
  - Timeliness,
  - Manipulation Resistance,
  - Martingale Property,
  - Verifiability,
  - Replicability,
  - Stability,
  - Parsimony.

✓ Empirical mean (Average Price):

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

See [Vin21]

Not robust to price manipulation

✓ Volume-Weighted Average Price (VWAP):

$$\widehat{\mathtt{VWAP}}_n := \frac{\sum_{i=1}^{n} V_i P_i}{\sum_{i=1}^{n} V_i}. \tag{2}$$

Used by [Nas22] to obtain the Nasdaq Crypto Index (NCI) and by FTSE Russell [FTS22] to compute Digital Asset Reference Prices.

Obeys to the FOREX symmetry rule

Not robust to price and volume outliers

✓ Volume-Weighted Median price (VWM):

$$\widehat{\mathtt{VWM}}_n := \inf\left\{p : \frac{\sum_{i=1}^n V_i \mathbf{1}_{P_i \le p}}{\sum_{i=1}^n V_i} \ge \frac{1}{2}\right\}. \tag{3}$$

See Federal Reserve Bank of New York [Fed15] on interest rates.

At work at the Chicago Mercantile Exchange [PK16] [CF 22b] [CF 22a] and Bloomberg [Gal18].

Robust to price outliers

Not robust to volume outliers

Issues/Questions

✓ high frequency update ⇒ small data set: $n = 100$

✓ any insight from theoretical control of statistical fluctuations?

✓ which method is the best? ⇒ **a new one = Robust Weighted Median**

An illustration on real data

Figure 1: Aggregated volume weighted price estimation on simulated data surrounding an efficient price generated every minute during one day using $\widehat{\mathtt{VWM}}_n$ (blue), $\widehat{\mathtt{VWAP}}_n$ (orange) and $\widehat{\mathtt{RWM}}_n$ (green).

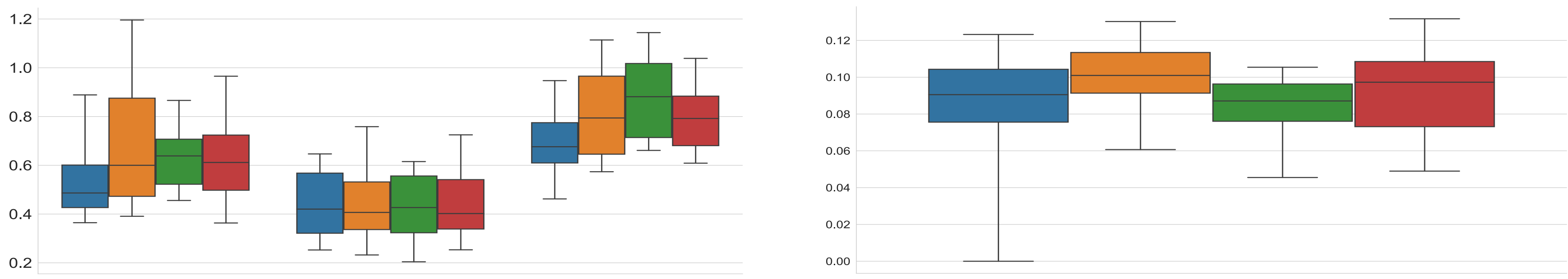
2 Understanding data characteristics

Using Kaiko's Instrument Explorer
https://instruments.kaiko.com/#/instruments and Trades product
https://docs.kaiko.com/#historical-trades.

- ✓ 67 pairs during 4 different days, including both stressed (2022-05-05, 2022-06-12) and calm (2022-06-28, 2022-12-09) periods in the market.
- ✓ 268 use cases (one pair for one period) for a total of 80,448,919 trades.

	Number of trades		Returns		
Period	Mean	Std.	Mean	Std.	Ann. Vol.
2022-05-05	96	156	$-0.44 \cdot 10^{-4}$	$8 \cdot 10^{-4}$	55%
2022-06-12	188	311	$-0.19 \cdot 10^{-4}$	$15 \cdot 10^{-4}$	104%
2022-06-28	94	149	$-0.24 \cdot 10^{-4}$	$8 \cdot 10^{-4}$	56%
2022-12-09	42	74	$-0.04 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	36%

Table 1: Data statistics (number of trades, returns) per minute by discarding 5% of the lowest and the highest returns.



(a) (b)

Figure 2: (a): Box plot of the estimated tail-index on the volumes (left), on the price returns (middle) and on the product of the two (right).
(b): Box plot of the estimated upper tail dependence.

Summary

- ✓ Both V and P are heavy tailed
- ✓ Dependence in the tail
- ✓ Need for robust aggregators
- ✓ Should satisfy the desirable properties of a price aggregator

3 Theoretical analysis of statistical fluctuations

Cumulative distribution function (c.d.f.) of the weighted price as

$$F_{\mathtt{W}}(x) := \frac{\mathbb{E}\left[W \cdot \mathbf{1}_{P \le x}\right]}{\mathbb{E}\left[W\right]}, \qquad x \ge 0. \tag{4}$$

$$\mathtt{WAP} := \frac{\mathbb{E}\left[W \cdot P\right]}{\mathbb{E}\left[W\right]}. \tag{WAP}$$

$$\mathtt{WM} := \arg\inf_{p} \ \mathbb{E}_{\mathtt{W}}\left[|P - p|\right]. \tag{WM}$$

$$q_{\mathtt{W}}(\alpha) := \inf\left\{p : \frac{\mathbb{E}\left[W \mathbf{1}_{P \le p}\right]}{\mathbb{E}\left[W\right]} \ge \alpha\right\}. \tag{$q_{\mathtt{W}}$}$$

Data-based definitions:

$$\widehat{\mathtt{WAP}}_n := \frac{\sum_{i=1}^n W_i P_i}{\sum_{i=1}^n W_i}, \tag{5}$$

$$\widehat{q_{\mathtt{W},n}}(\alpha) := \inf\left\{p : \frac{\sum_{i=1}^n W_i \mathbf{1}_{P_i \le p}}{\sum_{i=1}^n W_i} \ge \alpha\right\}, \tag{6}$$

$$\widehat{\mathtt{WM}}_n := \widehat{q_{\mathtt{W},n}}\left(\frac{1}{2}\right). \tag{7}$$

Assumptions:

(H0): The c.d.f. $F_{\mathtt{W}}$ in (4) has a density $f_{\mathtt{W}}$:

$$F_{\mathtt{W}}(x) = \int_0^x f_{\mathtt{W}}(x')\mathrm{d}x'. \tag{8}$$

Different tail assumptions:

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

($\mathbf{H}_{\mathbf{X}}^{\mathbf{\Gamma}}$): 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 $c > 0$.

Non-asymptotic fluctuations on WM and WAP(n fixed small)

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

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

Simplified bounds (Theorems)

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

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

4 Experiments on synthetic data

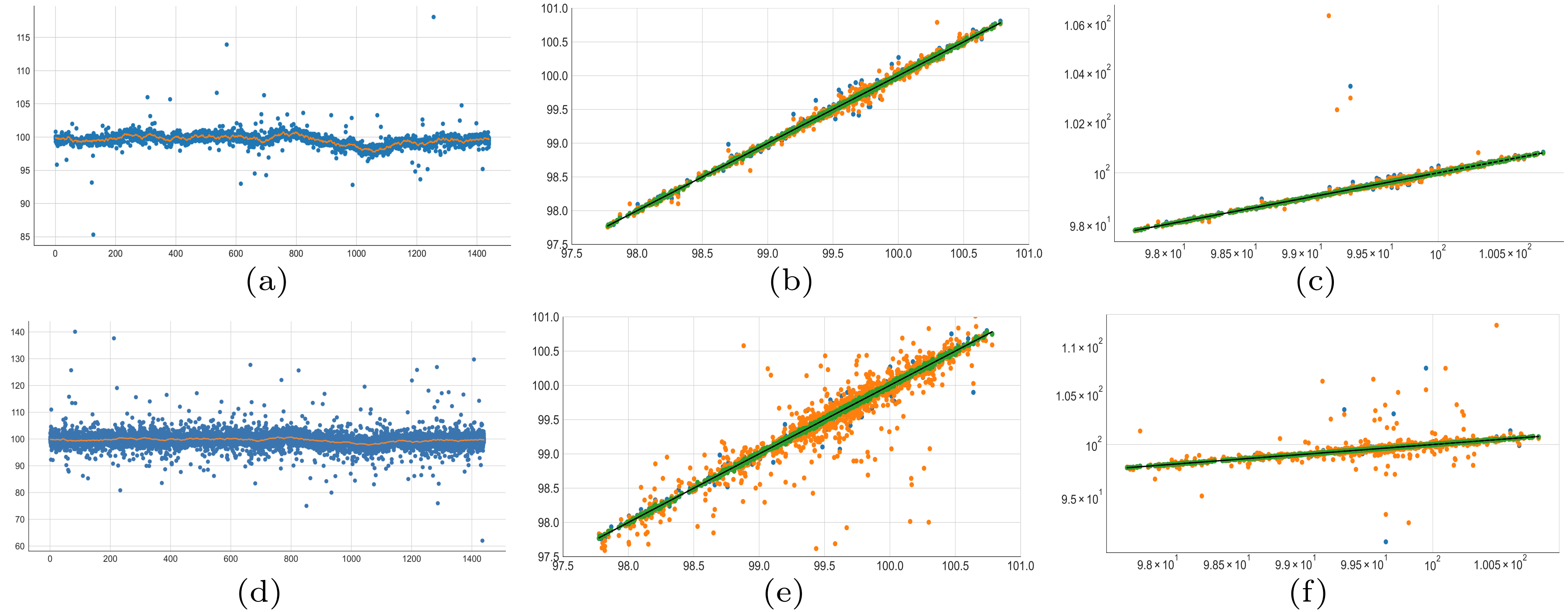


Figure 3: (a)-(d): Efficient price time-series $(S_t)_{t=1}^{1440}$ (orange line) simulated with $\sigma = 0.5$ ($\sigma^{\mathtt{eff}} = 0.51$), and its associated noisy prices $(\widetilde{S}_t^j, j \in \{1, \dots, 100\})_{t=1}^{1440}$ (blue dots) with $\omega = 0.99$ (a) and $\omega = 0.7$ (d).

(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. The best estimator is illustrated by the black dashed regression line $x \mapsto y = x$. Both horizontal and vertical axes are limited to 97.5 and 101 for a better display.

(c)-(f): Scatter plots of the reference price estimation with log-scale axes.

Computing the realized variance

	$\mathrm{RMSE}^{\texttt{price}}$			$\mathrm{RMSE}^{\texttt{RV}}$		
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 2: 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.

5 Experiments on real data

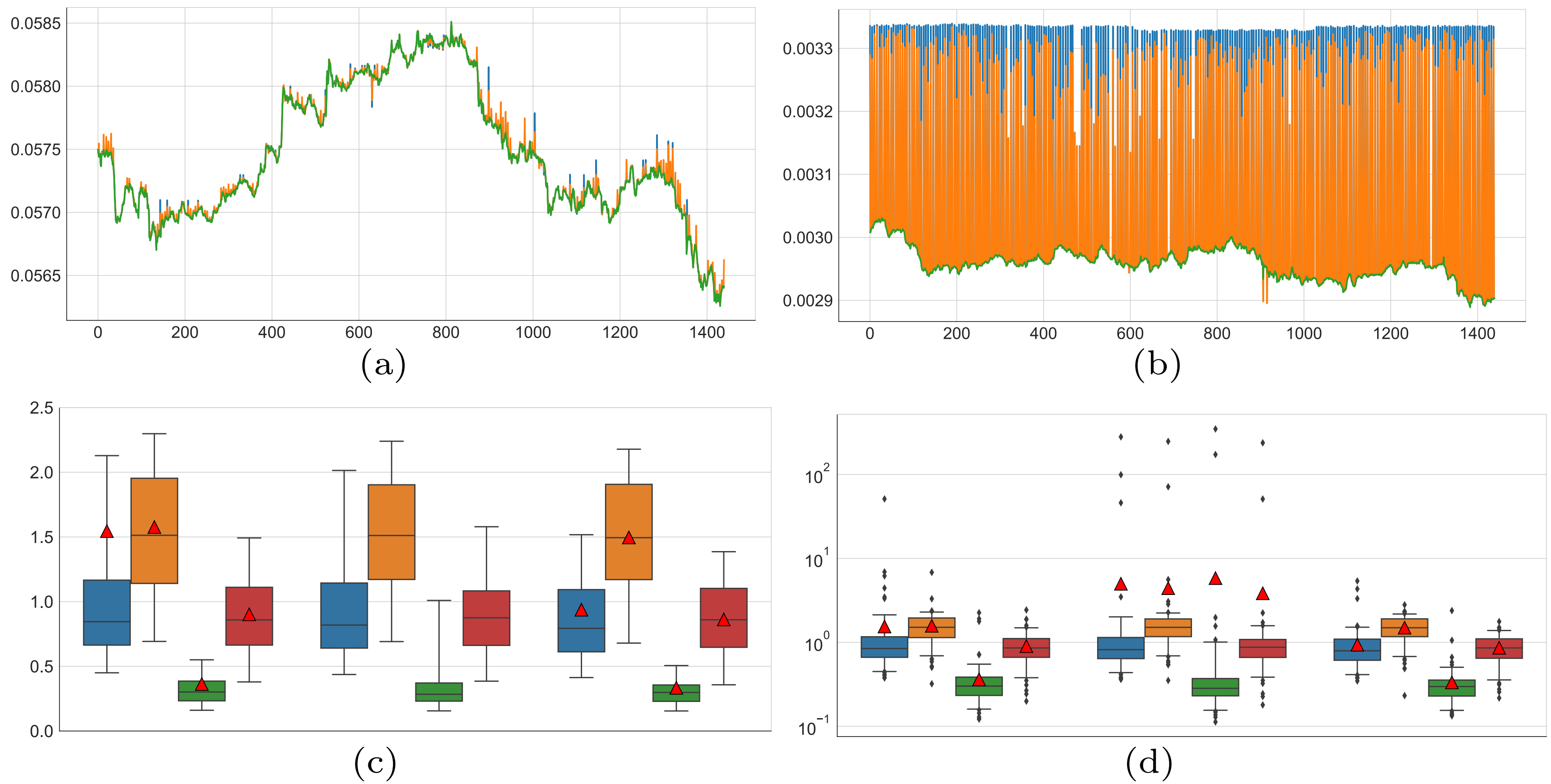


Figure 4: (a)-(b): Comparison between $\widehat{\texttt{VWM}}_n$ (blue), $\widehat{\texttt{VWAP}}_n$ (orange) and $\widehat{\texttt{RWM}}_n$ (green) on real data aggregation per minute of the pair eth/btc (a) and zec/btc (b) on 2022-06-28.
(c): Box plot of the annualized RV (over all pairs) from $\widehat{\texttt{VWM}}_n$ (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.

6 Conclusion

- ✓ Provide theoretical results on **non-asymptotic bounds** (n fixed) of the estimators.
- ✓ **CLT result** also available ($n \to +\infty$). Not interesting in practice.
- ✓ 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
- ✓ Preliminary step before modelling data with stochastic processes (multivariate, leadlag, asynchronous analyses etc)
- ✓ Extension done at Kaiko for cross-price of non-traded pairs

References

[AEMG22] M. Allouche, J. El Methni, and S. Girard. A refined Weissman estimator for extreme quantiles. *Extremes*, 2022.

[CF 22a] CF Benchmarks. CME CF cryptocurrency real time indices, version 15.1, December 2022. https://docs.cfbenchmarks.com/CME%20CF%20Real%20Time%20Indices%20Methodology.pdf.

[CF 22b] CF Benchmarks. CME CF cryptocurrency reference rates, version 15.1, December 2022. https://docs.cfbenchmarks.com/CME%20CF%20Reference%20Rates%20Methodology.pdf.

[Fed15] Federal Reserve Bank of New York. Technical note concerning the methodology for calculating the effective federal funds rate, July 2015. https://www.newyorkfed.org/medialibrary/media/markets/EFFR-technical-note-070815.pdf.

[FJS05] Gabriel Frahm, Markus Junker, and Rafael Schmidt. Estimating the tail-dependence coefficient: properties and pitfalls. *Insurance: mathematics and Economics*, 37(1):80–100, 2005.

[FTS22] FTSE Digital Asset Research. Guide to the calculation of the FTSE DAR Digital Asset Prices and FTSE DAR Reference Prices, November 2022. https://research.ftserussell.com/products/downloads/Guide_to_the_Calculation_of_FTSE_DAR_Digital_Asset_Prices_and_Reference_Prices_Fixes.pdf.

[Gal18] Colin Gallagher. CFIX methodology, Bloomberg cryptocurrency solutions, September 09 2018. https://data.bloomberglp.com/professional/sites/10/CFIX-Methodology.pdf.

[Nas22] Nasdaq. Nasdaq Crypto Index: index methodology, June 2022. https://indexes.nasdaqomx.com/docs/methodology_NCI.pdf.

[PK16] Andrew Paine and William J. Knottenbelt. Analysis of the CME CF Bitcoin Reference Rate and Real Time Index, October 20 2016. https://www.cmegroup.com/trading/files/bitcoin-white-paper.pdf.

[Vin21] Vinter. Crypto reference rates for single assets, version 2.1, February 2021. https://methodology.vinter.co/vinter/reference-rates.