Data review

Author: Vitalii Novikov

Data source

Kaggle: high-frequency-crypto-limit-order-book-data

General info

The dataset was created from raw order book messaged acquired by subscribing to the websocket of coinbase.com. The order books where then aggregated to construct snapshots of the limit order book at frequencies of 1 second, 1 minute and 5 minutes.

Content

The dataset contains roughly 12 days of limit order book data for Bitcoin (BTC), Ethereum (ETH) and Cardano (ADA). The data contains information for the 15 best bid / ask price levels in the order book.

Features

midpoint = the midpoint between the best bid and the best ask spread = the difference between the best bid and the best ask

{x} from 0 to 14 (number of levels: 15)

bids_distance_x = the distance of bid level x from the midprice in % asks_distance_x = the distance of ask level x from the midprice in %

bids_market_notional_x = volume of market orders at bid level x bids_limit_notional_x = volume of limit orders at bid level x bids_cancel_notional_x = volume of canceled orders at bid level x

asks_market_notional_x = volume of market orders at ask level x asks_limit_notional_x = volume of limit orders at ask level x asks_cancel_notional_x = volume of canceled orders at ask level x

EDA notes

Code and Data Dictionary available on Git: github.com/vitalii-novikov/Graph_Neural_Network_for_Market_Microstructure/tree/main/EDA_notebooks

Missing values: there are no missing values in any of the CSV files. **Zero values**: number of zero values is large (60%-90%) in *market_notional* and *cancel_notional* columns. In 1sec files the *limit_notional* columns includ a lot of zero values (~50%).

Notional volume

- usually the largest notional volume at the 0 level
- the distributions of volume at levels 1-14 are similar

- there are a lot of outliers at each level (sometimes the volume reaches inscane amounts, while usually is much lower)
- the most of orders executed via limit procedure (that's why *cancel* and *market* columns is usually represented by zero values)

Distance

- Negative values for bids distance (because below midpoint), and Positive values for asks distance
- 0 level is very close to midpoint (mean: -0.000011 (bid), mean: +0.000011 (ask), but even 25th percentile is 0.0)
- at levels 1-14 the absolute value of distance increases gradually

Financial correlation notes

Code available on Git: github.com/vitalii-novikov/Graph_Neural_Network_for_Market_Microstructure/tree/main

Metric	BTC ↔ ETH	ETH ↔ ADA	ADA ↔ BTC
Pearson correlation	0.7567	0.9353	0.8376
Lead-lag correlation (max lag)	0 steps (0 min)	0 steps (0 min)	0 steps (0 min)
*Return correlation *	0.8498	0.5757	0.5409
Full order imbalance correlation	0.2123	-0.0038	0.0081
Near order imbalance correlation	0.1242	0.0067	0.0495
Depth ratio correlation (near levels)	0.0531	-0.0024	0.0090
Depth ratio correlation (full levels)	0.0700	-0.0178	-0.0166

- **Strongest overall relationship:** between **ETH ↔ ADA** (Pearson 0.935), indicating very similar price behavior.
- **BTC ↔ ETH** also shows high co-movement (returns correlation 0.85), suggesting they often react together to market shifts.
- **ADA ↔ BTC** has moderate correlations (~0.54–0.84), implying partial but weaker synchronization.
- **Order imbalance correlations** are generally **low (0.00–0.21)**, showing that liquidity pressure is not strongly shared between markets.
- **Depth ratio correlations** are **close to zero or slightly negative**, suggesting **liquidity asymmetry** behaves independently across assets.
- **Lead-lag results** indicate **no significant time delay** (0 min shift), meaning none of the assets consistently leads or lags others at a 5-minute scale.