Return Predictions From Trade Flow

January 31, 2022

1 Introduction

Here you will assess trade flow as means of generating profit opportunities in 3 cryptotoken markets. We stress the word "opportunity" because at high data rates like these, and given the markets' price-time priority, it is far easier to identify desirable trades in the data stream than it is to inject oneself profitably into the fray.

2 Data

We have preprocessed level 2 exchange messages from the Coinbase WebSocket API for you into a more digestible format.

2.1 Treatment

Load the 2021 data for all 3 pairs from the class website. For each one, split it into test and training sets, with your training set containing the first 20% of the data and the test set containing the remainder.

2.2 Format

The data has the following structure[1]

2.2.1 Trades

received_utc_nanoseconds	timestamp_utc_nanoseconds	PriceMillionths	SizeBillionths	Side
1618090137140737000	1618090137157544000	35690	1000000	-1
1618090137851379000	1618090137864544000	35700	29801980	2
1618270615253262000	1618270615358639000	35760	2926932560	-1
1618270616012160000	1618270616105583000	35760	16673940	-1

The *Side* is actually a sum of trade sides at the same price and time.

2.2.2 Book

Ask1PriceMillionths	35700	35700	35770	35770
Bid1PriceMillionths	35690	35690	35760	35760
Ask1SizeBillionths	11872084060	11872084060	1255039420	1255039420
Bid1SizeBillionths	32957203990	32957203990	24752612680	24752612680
Ask2PriceMillionths	35710	35710	35780	35780
Bid2PriceMillionths	35680	35680	35750	35750
Ask2SizeBillionths	31032423370	30332423370	31011776970	31011776970
Bid2SizeBillionths	45284575470	45284575470	41785630850	41785630850
received_utc_nanoseconds	1618090136351018000	1618090136378911000	1618270617727565000	1618270617738680100
timestamp_utc_nanoseconds	1618090135799659000	1618090136388074000	1618270617836039000	1618270617846283000
Mid	35695	35695	35765	35765

(transposed)

[1]Note that inaccuracies in clock settings, i.e. "clock skew", can cause timestamps to appear later than the time at which they are recorded as having been received.

3 Exercise

Write code to find τ-interval trade flow $F_i^{(\tau)}$ just prior[2] to each trade data point[3] i. Compute T-second forward returns[4] $r_i^{(T)}$. Regress them against each other in your training set, to find a coefficient β of regression.

For each data point in your test set you already have $F_i^{(\tau)}$, so your return prediction is $\hat{r}_i := \beta \cdot F_i^{(\tau)}$. Define a threshold j for $\hat{r}_i$ and assume you might attempt to trade whenever $j < |\hat{r}_i|$.

4 Analysis

Assess the trading opportunities arising from using these return predictions in your test set. As part of this assessment, comment on the reliability of β, how you chose j, and what you might expect from using much longer training and test periods.

[2]We do not include the trade i data itself, because we are evaluating trade i in terms of the flow we would have been aware of just before it happened.

[3]NOTE: the trade data series does not necessarily have strictly increasing timestamps. Be sure not to include other trades at the same timestamp in your computation of F_i.

[4]It is not necessary to handle latency in your homework, but for your edification: a more careful implementation would account for lags. For a pessimistic approach we could choose L as, say, twice the 99th percentile of computational and communications lag. Then, it would use book data (not just trade data) to help compute return from time $t_i + L$ to $t_i + L + T$ and run regressions using that. The idea here is that it takes approximately time L to "do anything" about trade information.