Homework 1
Yunxiao Xiang – N12351630

Problem 0: Data Processing

In this part, I set mypath to where I store all the ".gz" files, and did the following:

1> Unzipped ".gz" files into same folder
2> Processed data into dataframe, with time converted into Timestamp object and set as index
3> Read data from each file and combine them in to two dataframes "quotes" and "trades"
4> Saved two dataframes as "trades.csv" "quotes.csv" files for future uses

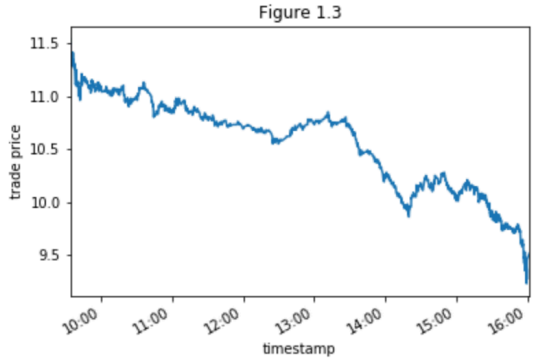
the head of the data frame looks like this:

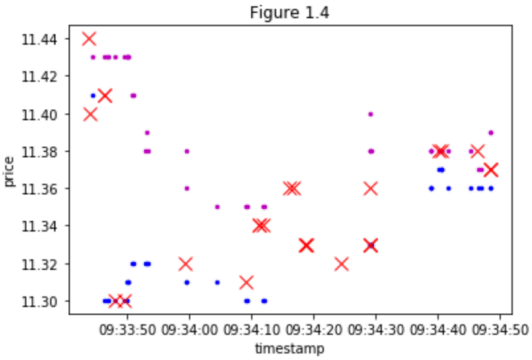
trades

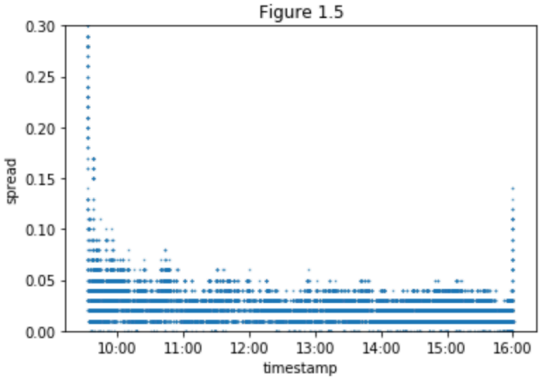
timestamp	symbology	price	size	exchange	saleconditions	iscorrection	iscancel	isopen	isclose
2008-10-01 09:30:43.844	CBL	19.89	100	P	@	false	false	false	false
2008-10-01 09:31:07.575	CBL	19.91	284	D	F	false	false	false	false
2008-10-01 09:34:11.675	CBL	19.74	200	D	F	false	false	false	false
2008-10-01 09:35:02.216	CBL	19.58	700	T	F	false	false	false	false
2008-10-01 09:35:02.216	CBL	19.58	700	T	Q	false	false	false	false
2008-10-01 09:35:05.816	CBL	19.58	1400	T	F	false	false	false	false
2008-10-01 09:35:07.556	CBL	19.58	100	T	F	false	false	false	false
2008-10-01 09:35:07.586	CBL	19.58	100	T	F	false	false	false	false
2008-10-01 09:35:08.096	CBL	19.58	100	P	F	false	false	false	false
2008-10-01 09:35:08.096	CBL	19.58	100	P	F	false	false	false	false
2008-10-01 09:35:08.126	CBL	19.58	200	T	F	false	false	false	false
2008-10-01 09:35:10.797	CBL	19.59	100	T	F	false	false	false	false
2008-10-01 09:35:11.097	CBL	19.4	100	T	@	false	false	false	false
2008-10-01 09:35:11.127	CBL	19.62	200	P	F	false	false	false	false
2008-10-01 09:35:11.127	CBL	19.62	100	P	F	false	false	false	false

Problem 1.1:

(1) In this problem, I reproduced Figures 1.3, 1.4, 1.5 according to trades and quotes in NYSE on 2008-10-15. Here are the plots:



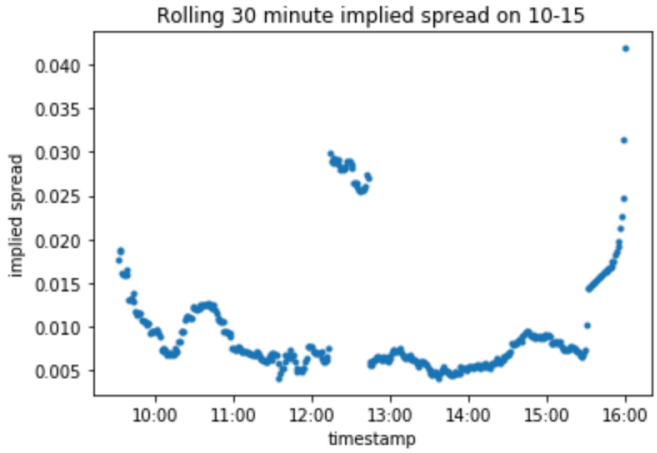




(2) And I compute the estimated first-order autocovariance of the price changes and the implied spread in the Roll model, by shift price to get difference in price ('dp') and shift dp to get autocov of dp. I'm using data on 2018-10-15 9:30 am – 16:00 pm, the computed

First-order autocov is: -6.505267086376985e-05
Implied spread is: 0.016131047190281213

(3) Then I plot the Roll implied spread using rolling 30-minute intervals, the plot looks like follows:



From the graph, the pattern roughly correspond to the smirk in 1.5 because the spread are higher at the beginning and the end of the day. What is different is the implied spread is high during 12-13, that might be caused by large price changes without actual spread changing too much.

Problem 2.1

Please see the last page

Problem 2.2

In this problem I implemented the quote and tick test to classify trades in NYSE, the classification algorithm is as follows:

- We first use quote test, where we look back into the previous second and find the nearest bid and ask price. If the price of the trade = last bid, it is classified as a “sell”. If the price of the trade = last ask, it is classified as a “buy”, if it is somewhere in the middle or we cannot find quotes in last second, we will raise exception and move to tick test
- In tick test we compare the current trade price with last price within 10 mins (I choose 10 mins because I think any price before 10 mins ago is not as meaningful), if the current price is higher, it is classified as a “buy”. If the current price if lower, it is classified as a “sell”. If the current price and last price are equal, we look for the next closest historical price and compare with the same logic. If there is no more price in 10min, it will classify the trade as “cannot be classified”.

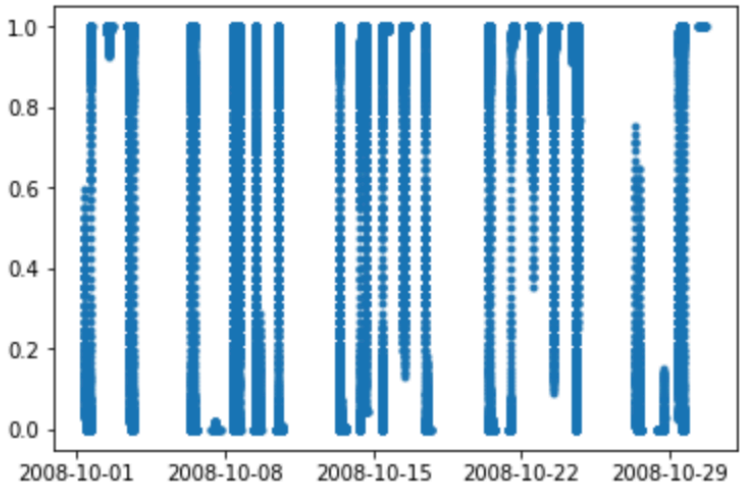
After I classified each trade, I add a column in the data frame called “Buy or sell” to store the type of the trade, and I also save the dataframe as "trades_with_classification.csv" file under “mypath”, the head of the data frame looks like this:

trades_with_classification

timestamp	symbology	price	size	exchange	saleconditions	iscorrection	iscancel	isopen	isclose	Buy or sell
2008-10-01 09:37:16.763	CBL	19.78	3000	N	@	false	false	false	false	Cannot be identified
2008-10-01 09:37:35.754	CBL	19.7	200	N	F	false	false	false	false	Sell
2008-10-01 09:37:41.455	CBL	19.75	100	N	@	false	false	false	false	Buy
2008-10-01 09:37:42.505	CBL	19.75	100	N	@	false	false	false	false	Buy
2008-10-01 09:37:44.845	CBL	19.72	200	N	@	false	false	false	false	Sell
2008-10-01 09:37:44.905	CBL	19.7	100	N	@	false	false	false	false	Sell
2008-10-01 09:37:44.905	CBL	19.7	200	N	@	false	false	false	false	Sell
2008-10-01 09:37:44.905	CBL	19.7	200	N	@	false	false	false	false	Sell
2008-10-01 09:37:51.565	CBL	19.78	100	N	@	false	false	false	false	Buy
2008-10-01 09:37:52.105	CBL	19.78	100	N	F	false	false	false	false	Buy
2008-10-01 09:38:22.287	CBL	19.71	100	N	F	false	false	false	false	Sell
2008-10-01 09:38:46.018	CBL	19.72	100	N	@	false	false	false	false	Buy
2008-10-01 09:38:46.018	CBL	19.72	100	N	@	false	false	false	false	Buy
2008-10-01 09:38:46.108	CBL	19.72	100	N	F	false	false	false	false	Buy
2008-10-01 09:38:58.289	CBL	19.65	200	N	F	false	false	false	false	Sell

Problem 2.3

In this problem we are using GM to compute the delta from the last observed delta, I create a new column in dataframe called “delta” to store the values, and the plot looks like:



Using class notation:

$$B = \mathbb{E}[V \mid \text{sell}]$$

$$= \underbrace{\mathbb{E}[V \mid \text{sell} \cap U] \cdot \mathbb{P}(U \mid \text{sell})}_{\text{I}} + \underbrace{\mathbb{E}[V \mid \text{sell} \cap I] \cdot \mathbb{P}(I \mid \text{sell})}_{\text{II}}$$

Part I: $\mathbb{P}(U \mid \text{sell}) = \dfrac{P(\text{sell} \mid U) \cdot P(U)}{P(\text{sell})}$

$$= \frac{\frac{1}{2}(1-\mu)}{\frac{1}{2}(2\delta\mu - \mu + 1)}$$

$$= \frac{1-\mu}{2\delta\mu - \mu + 1}$$

$$\mathbb{E}[V \mid \text{sell} \cap U] = (1-\delta)\bar{V} + \delta \underline{V}$$

Part II: $\mathbb{E}[V \mid \text{sell} \cap I] = \bar{V}$

$$\mathbb{P}(I \mid \text{sell}) = \frac{P(I \cap \text{sell})}{P(\text{sell})} = \frac{P(\text{sell} \mid I)\, P(I)}{P(\text{sell})}$$

$$= \frac{\mu \cdot 2\delta\mu}{2\delta\mu - \mu + 1}$$

$$\Rightarrow B = \bar{V} \cdot \frac{2\delta\mu}{2\delta\mu - \mu + 1} + \left[(1-\delta)\bar{V} + \delta \underline{V}\right] \frac{1-\mu}{2\delta\mu - \mu + 1}$$

$$= \frac{\bar{V}(\delta - 1)(\mu - 1) + \underline{V}\,\delta(\mu + 1)}{2\delta\mu - \mu + 1}$$