MATH-GA 2708.001 Algorithmic Trading and Quantitative Strategies

Homework 1: Market Impact & Optimal Trading

Profs. Petter Kolm and Lee Maclin
Due: March 28, 2025

Instructions

The following homework assignment can be done in groups of up to 3 students. Each group should turn in one write-up only and list the names and netIDs of all group members. Collaboration across different teams is not permitted. If someone in your team is not “carrying their load,” please let us know. We will be posting a separate document, "General Guidelines For All Homework" that will be more specific about deliverables and methodology.

Objective

In this assignment, you will:

- build a market impact model using the TAQ dataset,
- analyze the solution to a stochastic control problem, and
- examine standard concepts of statistical trading in the intraday / high-frequency domain.

Methodology and Deliverables

1. Building an Impact Model from Public Data.
   Background: Please refer to the lecture notes on impact models and the sample Python code that was shown in the presentation.

   Objective: We want to build an impact model, similar in form to the model described in Almgren et al.’s Direct Estimation of Equity Market Impact. Unlike their data set, our TAQ data set is public. We do not have trades classified as buyer or seller initiated, so we must infer direction using the tick test. We have to rely on the fact that classification errors will wash out during days dominated by noise trading – trading that is not driven by exogenous information.

List of stocks: We want to use a relatively liquid set of stocks to build our model, so we choose the most active stocks from our data set. You will get full credit by using the stocks from the S&P500. However, we strongly recommend you use more names than that; the largest 1500 stocks by average daily volume (that of course includes the S&P500) will give you better results. In addition to selecting from the SP500, after eliminating stocks that do not meet our criteria for contiguous data, you can use average daily value as a proxy for liquidity to select which stocks to add to your analysis.

Main data reading loop: Our main data reading and computation loop looks as follows:

```
For every stock in our list
    For every day of data
        Compute 2-minute mid-quote returns
        Compute total daily volume
        Compute arrival price – Average of first five mid-quote prices
        Compute imbalance between 9:30 and 3:30
        Compute volume-weighted average price between 9:30 and 3:30
        Compute volume-weighted average price between 9:30 and 4:00 (This will be used to compute average daily value of imbalance as described later in this document.)
        Compute terminal price at 4:00 – Average of last five mid-quote prices
```

When we are finished, we should have one matrix for each of the values we computed above. The dimensions of these matrices are: the columns are days and the rows are stocks.

From the matrix, we can eliminate the days (columns) in which any of our stocks have missing values. If a small number of stocks have many missing values, eliminate the rows corresponding to those stocks in the matrix rather than eliminating the columns of data in which most stocks have perfectly good values. Retain a record of which tickers you wind up using and which you discard.

Processing statistics: We will make one key departure from the model described in class. Almgren uses average daily volume traded, but we will use average daily value traded because this allows us to ignore splits. All else held equal, a two-to-one split means that the next day, twice as many shares will be traded at half of the previous price.

The daily value traded is defined as the volume weighted average price times the number of shares traded. The same is true of the imbalance, X. We will use the value of the imbalance, not the number of shares. The value of the imbalance is the number of shares multiplied by the volume-weighted price paid for those shares. Hence, if the imbalance is -400,000 shares, and the volume weighted average price calculated between 9:30 and 4:00 is $100, the daily imbalance value for that day is -$40,000,000.

In the end, we have one data point per stock per day. Each data point is comprised of an arrival price, a terminal price, a volume weighted average price, a volatility, an imbalance and an average daily volume. Using the regression equation provided in the lecture notes, we regress across all such data points to find eta and beta. (You will need to use NLS, a non-linear regression.)

As explained in the lecture notes, one of the assumptions in our impact model is that trade classification errors are not serially correlated, which is true on normal trading days, when there is little exogenous information and volatility is within some normal range. It may not be true on very volatile days. Hence, we want to run our analysis with and without those volatile trading days to see if it makes a difference in our estimates of eta and beta.

(a) **Testing your code** As explained in class, the data sets with which we are working are too large for the kind of visual error checking that normally takes place with smaller data sets. We cannot, for example, look at a series of graphs or scroll through a bunch of numbers in a spreadsheet and, from that information, determine that our code is working correctly.

Part of the objectives of this homework assignment is to show us that you have tested your code. You can do that by generating your own data using your regression model and then performing the regression to recover the parameters you used to generate that data. Lee described this process in class.

It is not necessary to do the above for the entire process, start to finish, but it is necessary to do it for each major step of the process. For example, suppose you have all of the matrices of data that you need to run your regression. One of those is a matrix of average daily volumes for each day for each stock. Other matrices represent other inputs to the model. Generate these matrices using the regression model as described in the previous paragraph. Then perform your regression on the generated matrices, not the real ones. Do you get back the parameters you used to generate these test matrices?

i. Deliverable: A set of unit tests that confirms your code works.

(b) **Perform non-linear regression for eta and beta:** The PDF of our class presentation shows how to back out permanent impact, leaving only temporary impact. We perform a non-linear regression (NLS) for our parameters. The regression should be performed across all stocks simultaneously (i.e cross-sectionally), not individually for each stock. Even so, there will be a tremendous amount of noise in our impact model.

i. Report the values and p-values (by using the two bootstrap techniques taught in class) of eta and beta: Please submit a text file, `params_part1.txt`, with the values of your eta and beta estimates. The format of this file is:
eta = [your value]
t-eta = [your value]
beta = [your value]
t-beta = [your value]
Are the parameters statistically significant?
Make sure to use *robust standard errors* in your calculations.

ii. Perform an analysis (as in Almgren et al.) of the residuals to determine their statistical properties. Do they satisfy the standard assumptions of nonlinear regression?

iii. Perform an analysis to determine whether eta and beta are different for the more active stocks in our data set as opposed to the less active stocks. Break the data set into two halves for this test.

iv. (Extra credit) Describe and perform a statistical test (e.g. White's general test for heteroskedasticity) to determine whether your residuals are homoskedastic or heteroskedastic.

2. Optimal Execution

(a) Consider the optimal liquidation problem with linear temporary and permanent price impact discussed in class.

i. Derive the solution of the resulting HJB equation. Then show that the optimal control and inventory process take the form

$$\begin{aligned} \nu_t^* &= -\sqrt{\frac{\phi}{k}}\,\frac{1+\zeta e^{2\gamma(T-t)}}{1-\zeta e^{2\gamma(T-t)}}\,Q_t^{\nu^*} \\ Q_t^{\nu^*} &= \frac{\zeta e^{\gamma(T-t)} - e^{-\gamma(T-t)}}{\zeta e^{\gamma T} - e^{-\gamma T}}\,q_0\,. \end{aligned}$$

Make sure to carefully show *all steps* in your answer.

ii. Compute the optimal control and inventory process when $\phi \to 0$. Provide an interpretation of the solution.

(b) Assume a stock's dynamic is $dS_t = \sigma dW_t$. You will determine the optimal liquidation trajectory of q_0 shares from t_0 to T using market orders subject to the value function

$$H(t_0, S, q) = \sup_{\nu \in \mathcal{A}_{t_0,T}} \mathbb{E}_{t_0,S,q}\left[\int_{t_0}^{T} (S_u - k\nu_u)\,\nu_u du + Q_T^{\nu}\,(S_T - \alpha Q_T^{\nu})\right]$$

where $k > 0$ is the temporary market impact, ν_u is the speed of trading, and $\alpha \geq 0$.

i. Derive the HJB equation for the value function and show that it satisfies

$$0 = -\left(\partial_q H - S\right)^2 - 4k\partial_t H - 2k\sigma^2 \partial_{SS} H\,.$$

ii. Derive the optimal liquidation rate ν_t^* as a function of the optimal inventory process $Q_t^{\nu^*}$ and problem parameters. Ansatz: $H(t, S, q) = h_2(t)q^2 + h_1(t)q + h_0(t) + qS$.

iii. Express the solution in (ii) as a function of q_0 and problem parameters.

iv. Determine the liquidation rate as $\alpha \to 0$ and provide an intuitive explanation for your result.

3. Standard Concepts of Statistical Trading.

(a) Describe four of the most commonly used high-frequency trading strategies. Support you claim of those being "most common" by providing some references. Distinguish between "alpha" strategies and other strategies.

(b) Provide a "back of the envelope" estimate of the profitability of high frequency traders in today's equity market. How do they use leverage? Motivate your assumptions, and how you come to your conclusion.

(c) Does high frequency trading impose risks of systemic nature? Find a journal paper or white paper *no older than two years*, that addresses this question. Do you agree or disagree with their findings?

(d) Propose your own intraday equity "alpha" trading strategy. Describe what methodology and data you would use to research your idea. Provide a "guestimate" of its performance (Sharpe ratio).

For (a) and (b) of the last question, you may find the following references useful:
http://www.sec.gov/news/press/2010/2010-8.htm
https://www.sec.gov/rules/concept/2010/34-61358.pdf
https://www.forbes.com/sites/simonconstable/2021/03/04/high-frequency-trading-helps-stabilize-the-stockmarket-new-research-shows/?sh=742622513500
https://seekingalpha.com/article/4319779-dhar-machine-learning-is-best-for-high-frequency-trading
http://www.cftc.gov/idc/groups/public/@economicanalysis/documents/file/oce_flashcrash0314.pdf