ATQS HW1: Processing TAQ Data

Yifan Li N14714308, Anna Zhang N10858751, Jingsheng Huang N14978082

February 2020

Abstract

In this assignment, we read and process the TAQ data and analyze the statistics of the cleaned data. Specifically, we work with stocks in the S&P 500 index and a time period from June 20, 2007 to September 20, 2007. Adjustment of data are based on corporate actions, and cleaning of data employed the method of a Bollinger band. In addition, we find the optimal bucket size to avoid the bid-ask bounce effect, so that the bucket returns display no autocorrelation. Lastly, we experiment with mean-variance optimization using the S&P 500 data and analyze the holdings of market portfolio.

1 Part A: Preparing the TAQ Data

1.1 Adjustments of Price and Shares Outstanding

The purpose of adjustment is ensuring that the impact of trades and quotes remains unchanged when corporate actions happen, such as stock splits. Price levels are adjusted by "Cumulative Factor to Adjust Prices" and "Cumulative Factor to Adjust Shares/Vol" columns in the "sp500.xlsx" file. The result of adjustment illustrates that not all stocks needed to adjusted. Essentially, we choose to take a further look at one stock (AAPL) that requires no adjustment and one stock (GILD) requires adjustment.

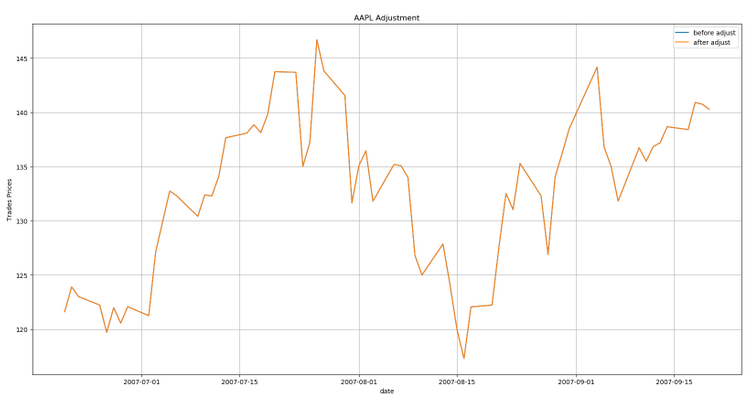

Figure 1: AAPL Adjustment

As expected, since no adjustment is required on Apple stock, the graph is showing two overlapping price levels. While for GILD stock, clearly there was stock splitting between June 22 and June 25, therefore adjustment executes and price levels are adjusted by Cumulative Factor to Adjust Prices (CFAP).

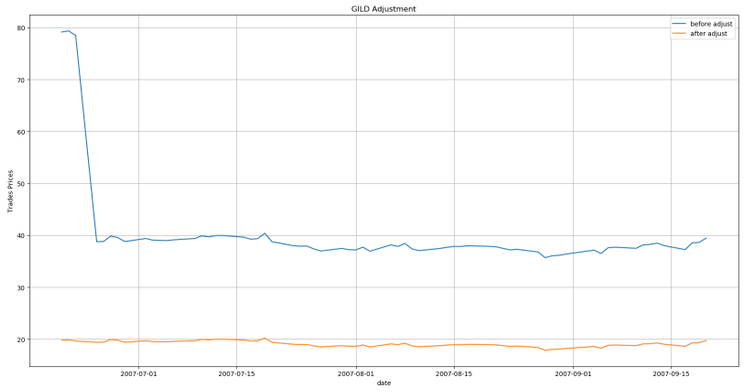

Figure 2: GILD Adjustment

1.2 Cleaning TAQ data

After adjusting price and volume levels, the next step is implementing cleaning procedure on TAQ adjusted data using the Brownlees and Gallo (2006) approach, where potential outliers are categorized using the following condition:

$$\left|\rho_i - \bar{\rho}_i(k)\right| < 2s_i(k) + \gamma \tag{1}$$

where are rolling window prices and volume standard deviation. After exploratory analysis on cleaned TAQ data, with same parameters, trades tend to have more outliers than quotes. The following figure illustrates cleaned result of UIS stock at August 10.

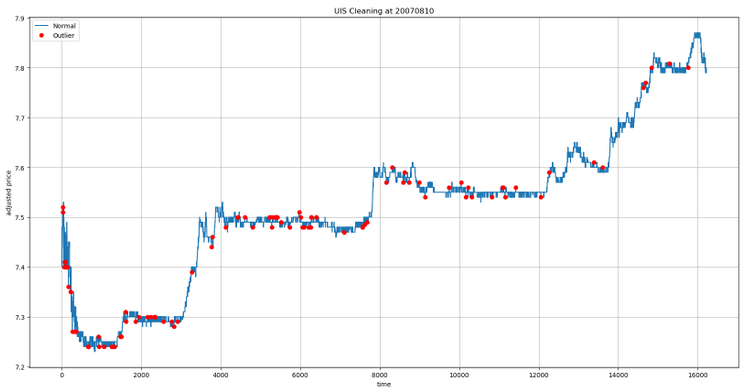

Figure 3: UIS trades cleaning

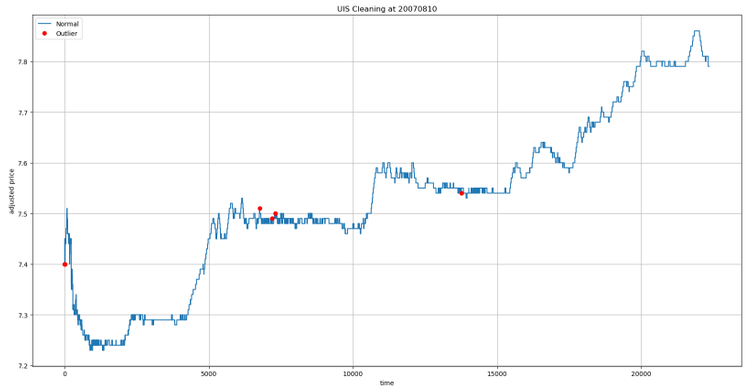

Figure 4: UIS quotes cleaning

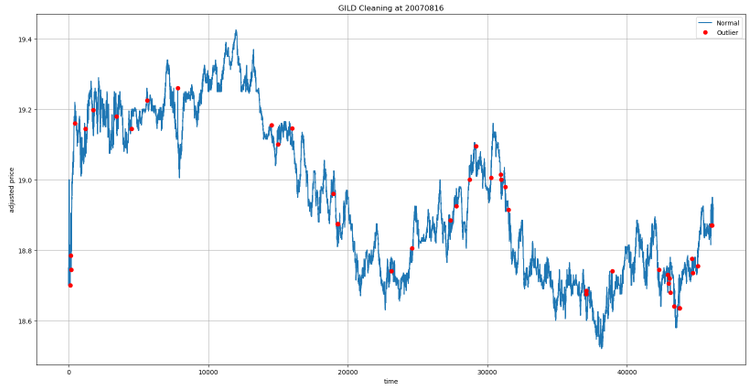

Figure 5: GILD trades cleaning

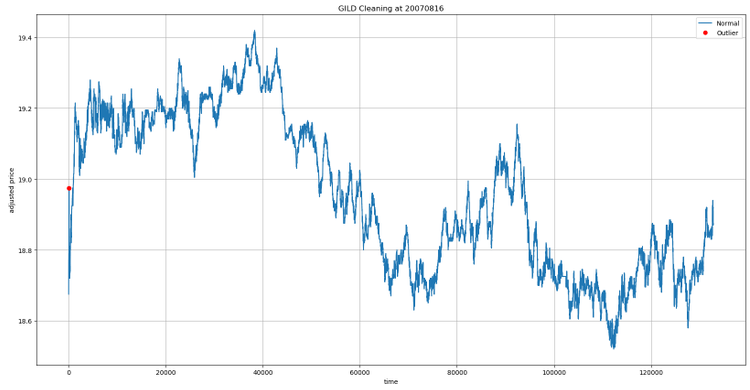

Figure 6: GILD quotes cleaning

2 Part B: Impact of Cleaning with Bollinger Band

In this section, we give the output of statistics for a few stocks and experiment with how different parameters of the Bollinger band might affect these statistics.

The following statistics are computed using a 5-minute interval, and $k = 25, \gamma = .0005$ in the Bollinger band.

1. GILD adjusted TAQ statistics before and after cleaning

   - Sample length in days for trades and quotes: (65, 65)

	before cleaning	after cleaning
# trades	1978898	1978344
# quotes	5411904	5411860
TQ ratio	0.3657	0.3656

Table 1: Basic trades and quotes information

	before cleaning	after cleaning
mean returns	(-0.16%, -0.29%)	(-0.13%, -0.25%)
median returns	(0.0%, 0.0%)	(0.0%, 0.0%)
standard dev.	(0.0352, 0.0354)	(0.0352, 0.0353)
MAD	(0.3806, 0.3818)	(0.3806, 0.3818)

Table 2: Annualized risk and return measures for (mid-quotes, trades)

	before cleaning	after cleaning
skew	(0.3563, 0.3456)	(0.3623, 0.3559)
kurtosis	(10.8325, 11.2862)	(10.8616, 11.3376)
max. drawdown	(-13.18%, -13.20%)	(-13.18%, -13.20%)

Table 3: Distributional measures and max. drawdown of returns

before cleaning	2.41	2.19	1.87	1.71	1.41	1.23	1.20	1.02	0.98	0.92
after cleaning	2.41	2.19	1.87	1.71	1.41	1.23	1.20	1.02	0.98	0.92

Table 4: 10 largest returns (%) for mid-quotes

before cleaning	2.44	2.19	1.86	1.71	1.56	1.30	1.20	1.06	0.99	0.95
after cleaning	2.44	2.19	1.86	1.71	1.56	1.30	1.20	1.06	0.99	0.95

Table 5: 10 largest returns (%) for trades

before cleaning	-1.51	-1.46	-1.41	-1.41	-1.25	-1.16	-1.16	-1.10	-1.08	-1.05
after cleaning	-1.51	-1.46	-1.41	-1.41	-1.25	-1.16	-1.16	-1.10	-1.08	-1.05

Table 6: 10 smallest returns (%) for mid-quotes

before cleaning	-1.57	-1.53	-1.50	-1.44	-1.25	-1.25	-1.19	-1.12	-1.11	-1.07
after cleaning	-1.57	-1.53	-1.50	-1.44	-1.25	-1.25	-1.19	-1.12	-1.11	-1.07

Table 7: 10 smallest returns (%) for trades

2. UIS adjusted TAQ statistics before and after cleaning

   - Sample length in days for trades and quotes: (65, 65)

	before cleaning	after cleaning
# trades	559021	556551
# quotes	1177957	1177887
TQ ratio	0.4746	0.4725

Table 8: Basic trades and quotes information

	before cleaning	after cleaning
mean returns	(-0.70%, -0.83%)	(-0.67%, -0.91%)
median returns	(0.0%, 0.0%)	(0.0%, 0.0%)
standard dev.	(0.0379, 0.0399)	(0.0376, 0.0384)
MAD	(0.4052, 0.4493)	(0.4052, 0.4490)

Table 9: Annualized risk and return measures for (mid-quotes, trades)

	before cleaning	after cleaning
skew	(0.1566, -0.6448)	(0.3172, -0.0053)
kurtosis	(11.2772, 22.5661)	(10.1262, 9.6551)
max. drawdown	(-32.83%, -32.75%)	(-32.83%, -32.75%)

Table 10: Distributional measures and max. drawdown of returns

before cleaning	2.07	1.97	1.73	1.68	1.64	1.51	1.46	1.29	1.27	1.26
after cleaning	2.07	1.97	1.73	1.68	1.64	1.51	1.46	1.29	1.27	1.26

Table 11: 10 largest returns (%) for mid-quotes

before cleaning	2.15	2.10	1.93	1.68	1.51	1.46	1.36	1.33	1.31	1.21
after cleaning	2.10	1.93	1.68	1.51	1.46	1.36	1.31	1.21	1.19	1.13

Table 12: 10 largest returns (%) for trades

before cleaning	-2.29	-1.90	-1.70	-1.58	-1.47	-1.43	-1.39	-1.34	-1.28	-1.24
after cleaning	-1.90	-1.70	-1.58	-1.47	-1.43	-1.39	-1.34	-1.28	-1.24	-1.15

Table 13: 10 smallest returns (%) for mid-quotes

before cleaning	-4.16	-2.03	-1.64	-1.58	-1.47	-1.45	-1.39	-1.34	-1.25	-1.25
after cleaning	-2.49	-2.03	-1.64	-1.39	-1.34	-1.34	-1.25	-1.25	-1.19	-1.11

Table 14: 10 smallest returns (%) for trades

We repeat the above procedure on GILD data for the bucket size of X = 10s, 30s, 1-minute, 5-minute, 10-minute, and 30-minute to compare the statistics. Please refer to the file ÍmpactAnalysis.pyfor output.

Furthermore, we experiment with different k and gamma in the Bollinger band to see their impact on the statistics. Please refer to the code zip file for output.

3 Part C: Ljung-Box Test for Serial Correlation

In this part of the problem, We find the optimal bucket size to compute returns. We define "optimal" to be the time interval such that the returns of both mid-quote and trade data display no serial correlation.

We first find optimal bucket for the ticker GILD to be the 11-minute interval, using mid-quote and trade returns between 2007-06-20 and 2007-07-04. The input bucket parameter ranges from 1 to 60 with a step-size of 5. The table below shows p-values at the 11-minute interval for a lag from 1 to 10; this is the smallest bucket size for which the p-values are all greater than 0.05. The null hypothesis in Ljung-Box test is that the data between each bucket exhibit no serial correlation, and a p-value bigger than 0.05 indicates statistical insignificance, thus the null hypothesis cannot be rejected.

lags	trade p-values	mid-quote p-values
1	0.08891251	0.07943525
3	0.19735502	0.16549501
5	0.28853853	0.22892995
8	0.60152759	0.52782662
10	0.72720194	0.65204322

Table 15: p-values for GILD data

Using IBM's data within the same time frame, the optimal bucket size is found to be the 6-minute interval. Experimenting with different tickers, we see that the optimal bucket size can be different for each ticker. All the computations use a lag parameter that ranges from 1 to 10.

lags	trade p-values	mid-quote p-values
1	0.93744528	0.90736457
3	0.558409	0.64136452
5	0.63012298	0.73615376
8	0.47603449	0.61756488
10	0.48831538	0.66713157

Table 16: p-values for IBM data

4 Part D: Analysis of Mean-Variance Optimization

The mathematical description of the problem being solved is:

$$\begin{array}{rl} \text{minimize} & -\bar{p}^T x + \mu x^T \Sigma x \\ \text{subject to} & 1^T x = 1, x \geq 0 \end{array}$$

The given code solves a portfolio optimization problem with 4 assets, where the mean return is $\bar{p} = [.12, .10, .07, .03]^T$ and the covariance matrix is

$$S = \begin{bmatrix} 0.04 & .006 & -.004 & 0.0 \\ .006 & .01 & 0.0 & 0.0 \\ -.004 & 0.0 & .0025 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

From the covariance matrix, we can infer that asset 4 is a risk-free asset with a 3% risk-free return. As the returns decrease from asset 1 to asset 3, the volatilities (standard deviations of returns) decrease as well. The correlation between assets are $\rho_{12} = 0.3, \rho_{13} = -0.4, \rho_{23} = 0$.

The output figure shows the optimal risk-return trade-off curve. The left-end point denotes 100% allocation in risk-free asset (corresponds to zero std.) The right-end point denotes 100% allocation in asset 1 which has the highest mean return. The lower graph shows the optimal asset allocation vector $\mathbf{x}$ (i.e. the weights). The result indicates for small risk, the optimal allocation would be mostly in the risk-free asset, a more aggressive portfolio would consist more of asset 1 and asset 2, which is what we expected. The optimization converges with absolute tolerance 10^{-7} and relative tolerance 10^{-6}.

4.1 Holdings of the Market Portfolio

We calculate the holdings of the market portfolio on June 20, 2007 and September 20, 2007 as

$$\frac{\text{adjusted price for stock } i \times \text{adjusted shares outstanding}}{\sum_{i=1}^{n} \text{adjusted price for stock } i \times \text{adjusted shares outstanding}}$$

where n is the total number of stocks in the market portfolio (i.e. using S&P 500 as proxy). The tables below show the top five holdings on each of these dates:

Ticker	Company Name	Holdings
XOM	Exxon Mobil	3.69%
GE	General Electric	3.18%
MSFT	Microsoft	2.27%
C	Citi	2.09%
T	AT&T	1.93%

Table 17: Top 5 holdings in market portfolio on June 20, 2007

Ticker	Company Name	Holdings
XOM	Exxon Mobil	4.01%
GE	General Electric	3.32%
MSFT	Microsoft	2.10%
C	Citi	2.01%
T	AT&T	1.84%

Table 18: Top 5 holdings in market portfolio on September 20, 2007

For question 3 in part D, we introduce the same way in question 2 to make convex optimization on S&P500 stock data. To do the optimization, we choose the period from 20070620 to 20070920 and filter out the stocks with complete return data in the whole period. We finally get 500 stocks. Then we calculate the daily excess return for each stock and make then into a matrix, rows of which stand for dates and columns of which stand for stocks. It is a 65*300 matrix. We can easily get the covariance matrix and a vector of total returns in this period as matrix sigma and vector p in problem, then with a list of different penalizing parameters mu we use cvxopt module to optimize the problem and get the optimal weights under each penalizing parameter.

we give two extreme examples as follows:

when mu close to 0, we filter out all weights higher than 0.0001 and only get one point whose value is very close to 1, which means, when we do not care about the volatility, we just all buy the stock with the highest expected return.

when mu is extremely high, we get the second picture in which the highest weight of a stock in the whole portfolio is still less than 10%, indicating a fully diversified portfolio.

the following graph shows the relationship between turnover ratio and the penalizing parameter with which the portfolio starts. For example, the curve starts with turnover rate 1, which means when starting with the portfolio gotten from the optimizer with penalizing parameter close to 0, we almost change all holding stocks during the period. That makes sense because we start with one stock and end with a fully diversified portfolio. And when starting with higher penalizing parameters, we will get lower turnover rates.

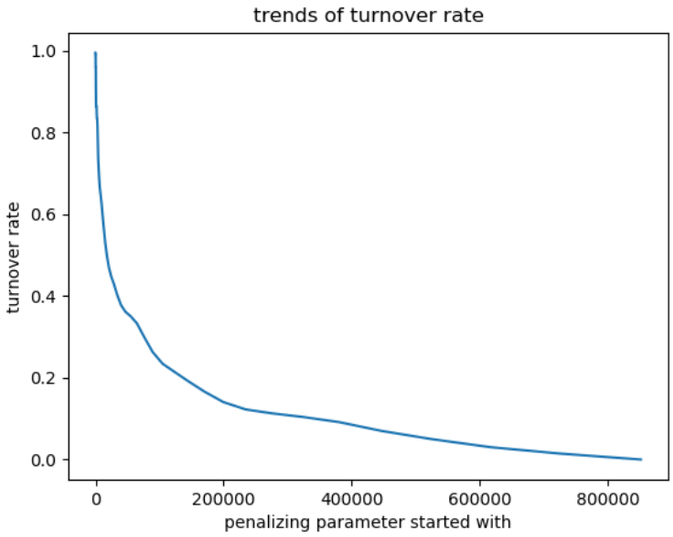

Figure 7: Trends of turnover rate

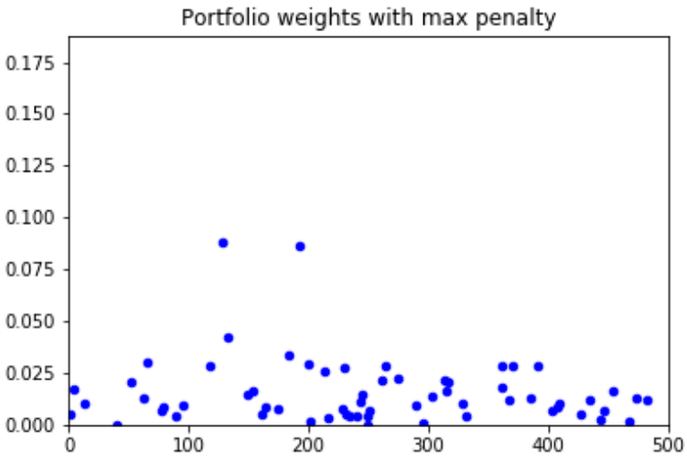

Figure 8: Portfolio Weights with Max. Penalty

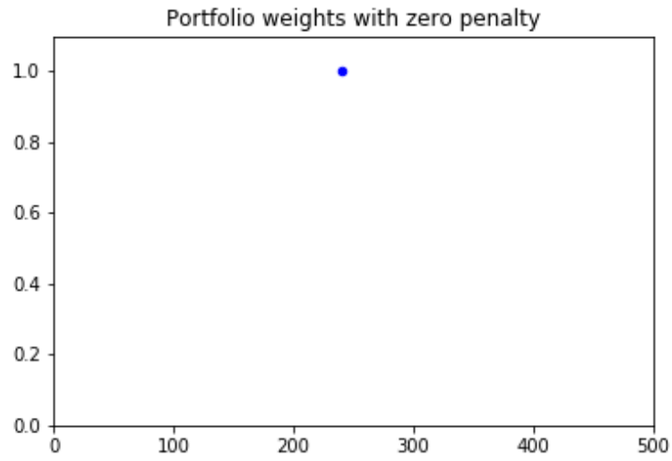

Figure 9: Portfolio Weights with Zero Penalty