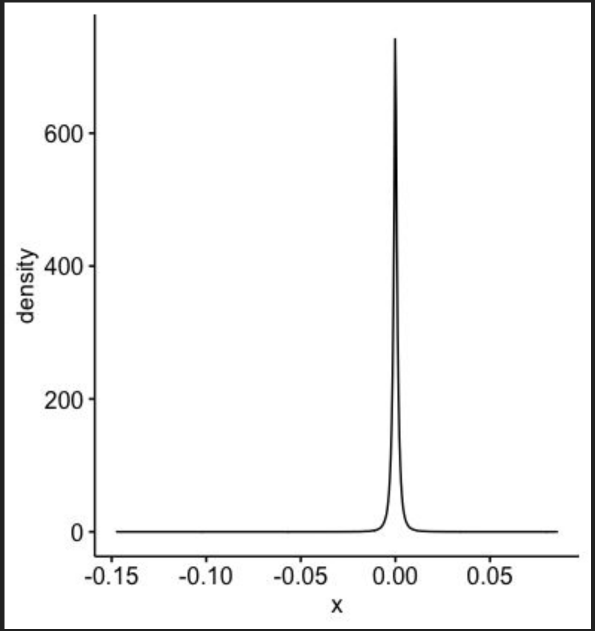
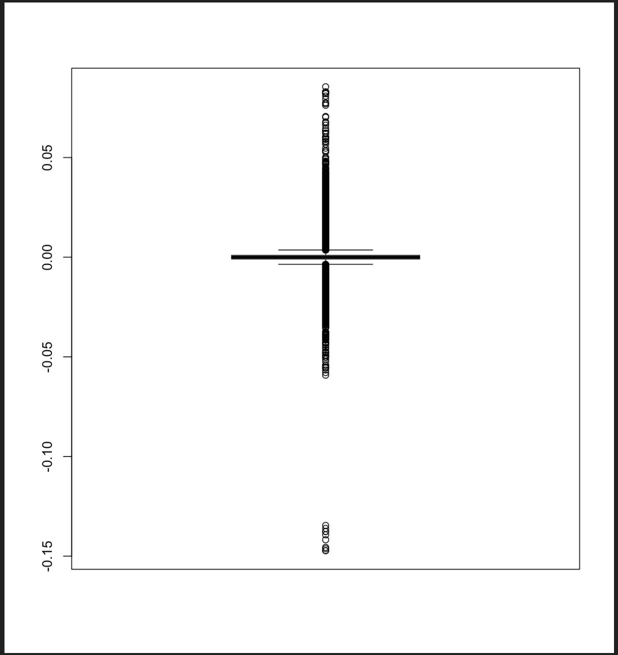
415 Final Project Advanced

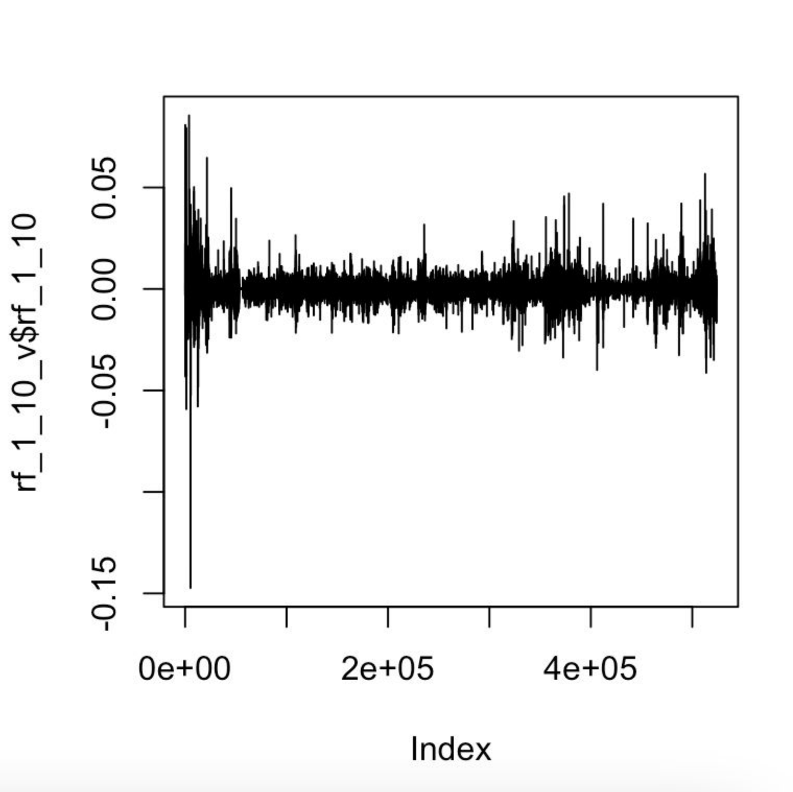
Yanyijie Zhou, Haijian Wu, Rui Qiu

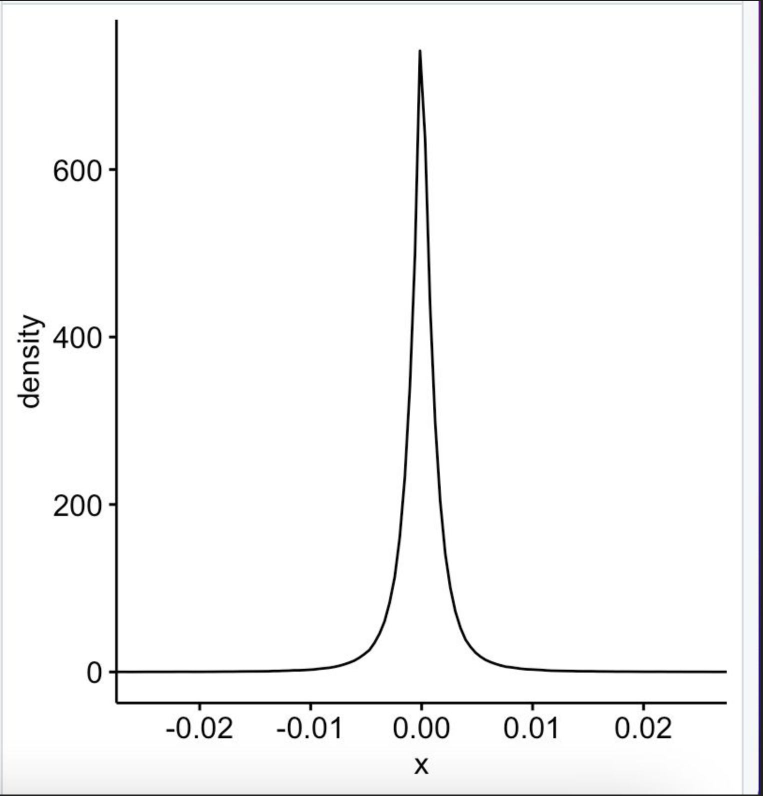
What we found about the given price dataset

- Real Dataset
- Outliers
- Noise






rf_1_10_v$rf_1_10
0.05
0.00
-0.05
-0.15
0e+00
2e+05
4e+05
Index


density
600
400
200
0
-0.02
-0.01
0.00
0.01
0.02
x

Features we tried and used in the end

- Features we tried
  - Backward Return
  - Forward Return
  - Asset Price
  - ARIMA features
  - Moving average & weighted moving average
  - Stock Price Indicators: (Abandoned due to computational complexity)
    - Relative Strength Index (comparison of losses and recent gains / overbought/ oversold)
    - Accumulation/Distribution Oscillator (indicator of momentum, using Highest - Lowest)
    - Average True Range (Measuring bull and bear price trends)
- Features we used in the end
  - Backward returns of asset 1-3 (chosen by AIC)

Model training & evaluation

1. Tried lots of different models
    a. OLS
    b. KNN
    c. ARIMA (Popular model in stock price prediction)
    d. Simplified Auto Regression
    e. Random Forest
    f. SVR (regression version of SVM)
    g. Gradient Boosting
2. Notice that these models perform worse on OJ -- overfitting or other issues?
3. Not applicable on OJ / Perform too slow
    a. ARIMA / KNN & Random Forest & SVR
4. Look through dataset -- outliers & noise

Model training & evaluation

1. Data Cleaning
   a. Remove the approximate values
   b. Remove the extreme values
2. Model Training
   a. Different models
   b. trunc(0.8 * nrow(data)) & nrow(data)
3. Model Selecting (based on AIC)
   a. AIC -- useful in time series prediction
   b. BIC -- confident that variables we use include the “real” variables
   c. adjr2 -- for training data (observations)

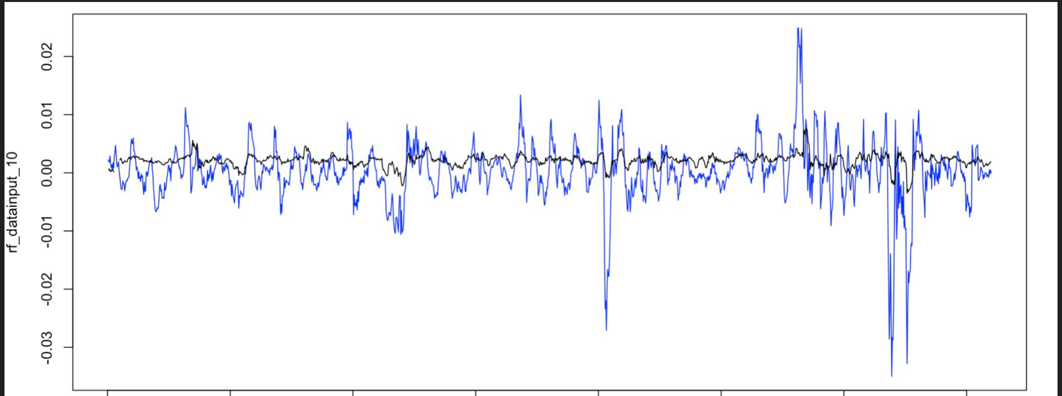
Model training & evaluation

1. Local Test

```r
corr <- cor(pred, real)
print(corr)

# RECORDs 1000 times
# OLS (lasso): -0.150
# OLS (RB+RF): -0.071
# OLS (SVM): -0.105
# OLS (RB + 1440): -0.047
# OLS (RB + 1440+asset): -0.043
# OLS (clean + aic + RB(1440)): 0.073
# Boost: 0.14 -> 0.098 -> 0.0689 -> 0.139 -> 0.110(aic)
# ols aic: -0.061
```

```r
rm(list=ls())
source("prediction.R")

originDATA <- read.csv("final_project_data.csv")
real_rf_10 <-read.csv("rf_10.csv")
originDATA <- originDATA[, c(2,3,4)]

h = 10
window = 1440
trytimes <- 1000
pred = rep(0, trytimes)
real = real_rf_10[(nrow(originDATA)  - h - trytimes + 1):(nrow(originDATA)  - h), 1]
real_check = rep(0, trytimes)

for(i in (nrow(originDATA) - window - h + 1 - trytimes + 1):(nrow(originDATA) - window - h + 1)){
  index  = i - (nrow(originDATA) - window - h + 1 - trytimes + 1) + 1;
  datainput <- originDATA[i:(i+window - 1),]
  pred[index] <- prediction(datainput)
  # real_check[index] <-  real_rf_10[(i+window - 1), 1]
}

corr <- cor(pred, real)
print(corr)
```

2. Visualizing



Accelerating Our Code

1. Model Fitting & Local Testing
    a. **Model Selecting**: Abandoned most non-parameter models (KNN & Random forest & SNR).
    b. **Predictors Selecting**: Abandoned all computational demanding features
    c. **Local Testing**: Perform parallel computing.
    d. **Overfitting Problem**: Remove unnecessary features from our model
    e. Using Python or Using Rccp to include C code in R may be helpful
2. Prediction
    a. Optimize the process of getting input / output value
        i. **Input**: Deleted all for-loops - O(N) v.s. O(Nn)
        ii. **Output**: OLS predicted value = predictor vector %*% coefficient vector
3. Prettified Codes & Detailed Comments

What we can do to improve

1. Trying more features & models
    a. The final version of our model was decided one day before the submission deadline. Some possible predictors & models were abandoned due to limited submission times.
2. More organized version control
    a. Forgot to record every submission & error feedback
    Spent too much time on debugging

Thank you

Reference:

- https://www.sciencedirect.com/science/article/pii/S2405918818300060
- https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2
- https://otexts.com/fpp2/