High Frequency Statistical Arbitrage Model

Pair trading using price movement per second in correlated companies

Dottie, Luisa, Cedrick, Tyler, & Vidushi

MS&E 448

Stanford University | Spring 2019

Outline

- Review since midterm
- Model selection
- Simulation
- Results
- Discussion and future work

Outline

- Review since midterm
- Model selection
- Simulation
- Results
- Discussion and future work

Review

Midterm milestone:

- *Company selection*
    - Two methods discussed:
        - Spherical KMeans on the whole features vectors
        - Euclidean KMeans at every time stamp
- *Cointegration of clusters:*
    - Chose to focus on just pairs

Review

Defining the universe:

- *Final train/test set:*
    - Train: one day of data 1/28/19 from 9:30am-4pm
    - Test: following three days

- *Final company selection:*
    - Two types of results:
        - Clustering
        - 19 most correlated pairs, able to execute trades on 14

Outline

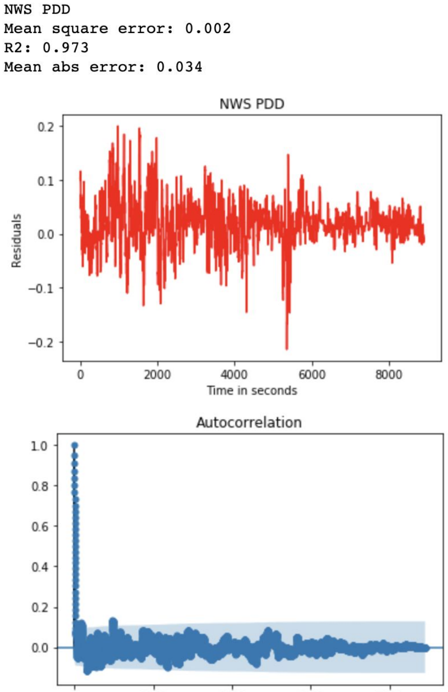
- Review since midterm
- Model selection
- Simulation
- Results
- Discussion and future work

Building the models

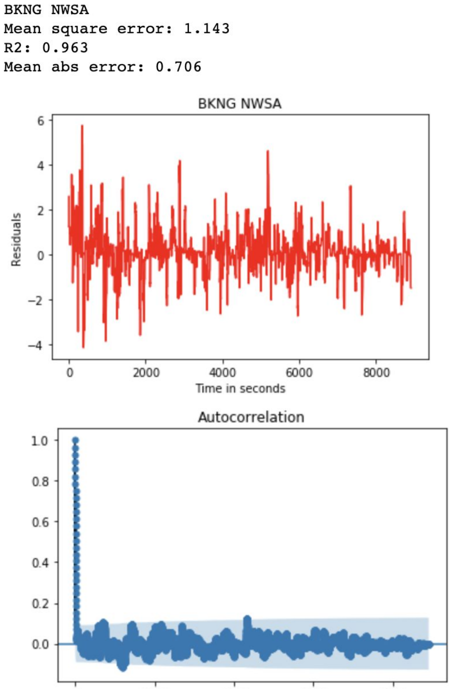
- Split train set into train and validation sets
- Add features: lags of `wmid` and `volume` at 15 and 30 seconds
- Run LASSO, forward stepwise, backward stepwise, and both ways on all features
- Compare methods using MSE

Features: `bid`, `ask`, `mid`, `wmid`, `bsize`, `asize`, `anum`, `bnum`, `volume`, `notional`, `last_price`, `last_size` + `lags`

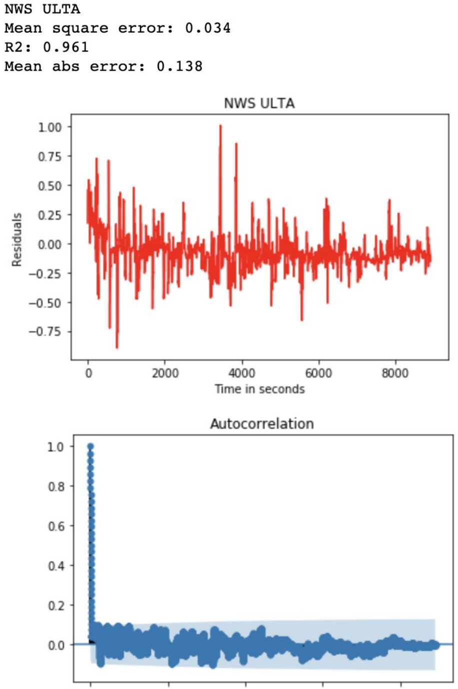
Testing the models



```
NWS PDD
Mean square error: 0.002
R2: 0.973
Mean abs error: 0.034
```



```
BKNG NWSA
Mean square error: 1.143
R2: 0.963
Mean abs error: 0.706
```



```
NWS ULTA
Mean square error: 0.034
R2: 0.961
Mean abs error: 0.138
```

**OOS test from 9:30 am - 12 pm*

Outline

- Review since midterm
- Model selection
- Simulation
- Results
- Discussion and future work

Simulation Strategy

- Best model for each pair selected
- Start with $10,000, 1 day after training period
- Model used to predict residual return 30 seconds in future
- If prediction is falls outside `num_std` of residual return, execute a trade
  - Band selection: num_`std` is a hyperparameter
  - Quantity limited to 1, executed on midprice instead of buy/ask
- Model trained in *online* fashion, i.e. model trained as more data is received

Backend and testing details

- Simulation based on Thesys Technologies simulator
- Difficult to work with simulator
  - Lack of API reference documentation
  - Uptime on server is poor, outages prevented us from testing
  - Missing data needs to be taken into account at runtime
- Performed runs over various timescales, i.e. minutes, hours, weeks after training period
- Performed runs with different bands to optimize `num_std`

Outline

- Review since midterm
- Model selection
- Simulation
- Results
- Discussion and future work

Results: Pair Trade NWS and ULTA

NWS

Company Name: News Corp

About: Mass Media and Publishing company

Industry: Media

Current Stock Price: $11.58

ULTA

Company Name: ULTA Beauty

About: Beauty Products, Makeup

Industry: Cosmetics

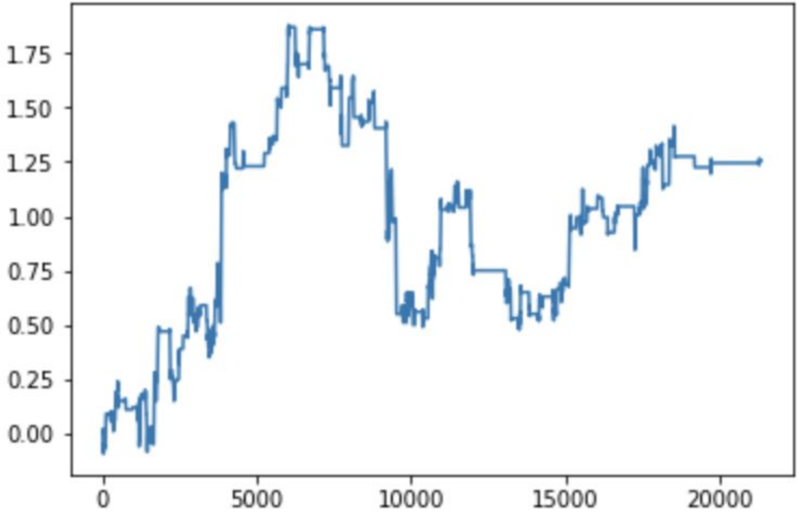
Current Stock Price: $335.53

TAKEAWAYS:

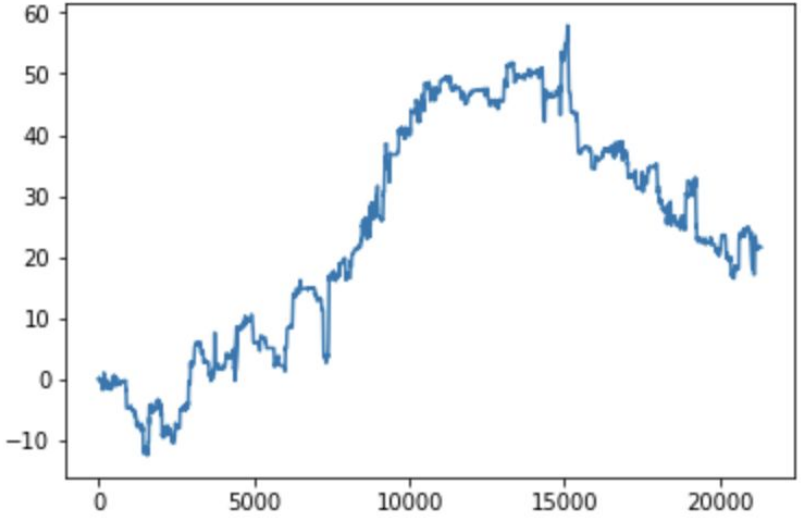
- Worried about spurious correlation
- Model has strong signals and high prediction capabilities

Results: Profit/Loss 3 Days (10:00am-12:00pm)

NWS: $1.26

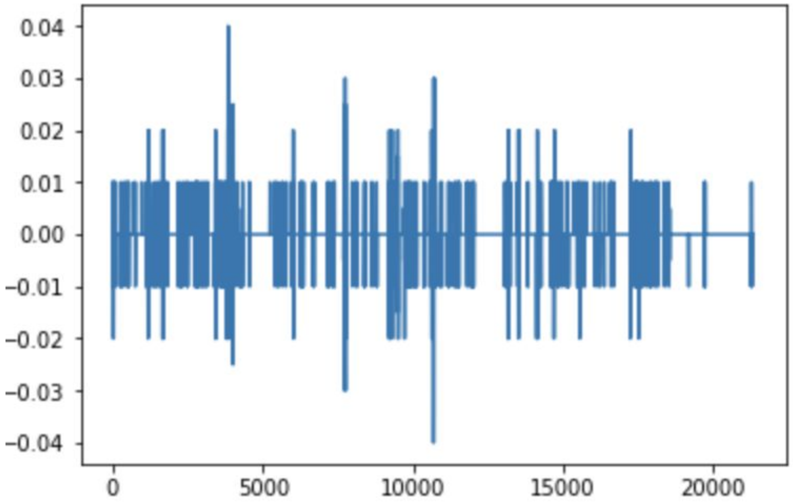


ULTA: $21.60

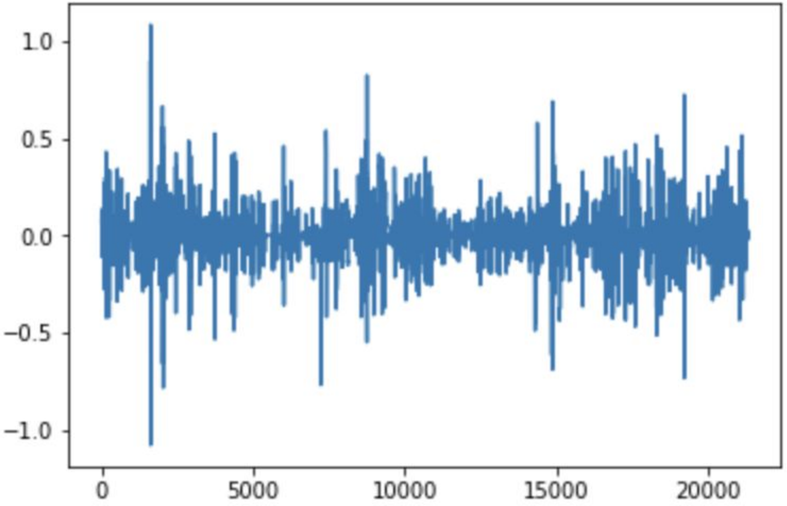


Results: Returns 3 Days

NWS

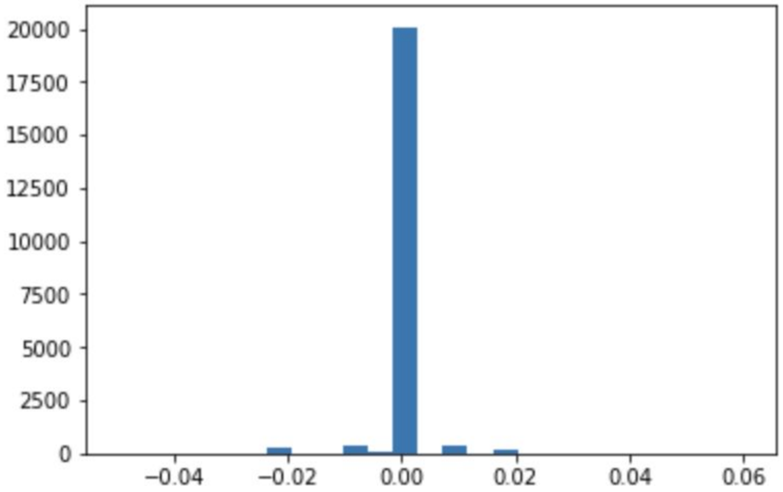


ULTA

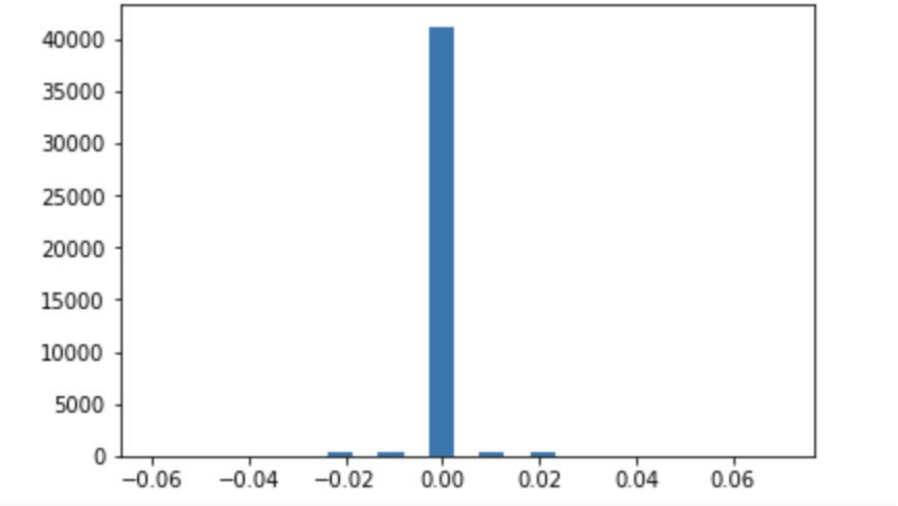


Results: Histogram of Returns 3 Days

NWS



ULTA



Results: Pair Trade VRSK and VIA

VRSK

Company Name: Verisk Analytics

About: Data Analytics and Risk Management

Industry: Data Analytics

Current Stock Price: $141.84

VIA

Company Name: Viacom

About: Mass media conglomerate

Industry: Media

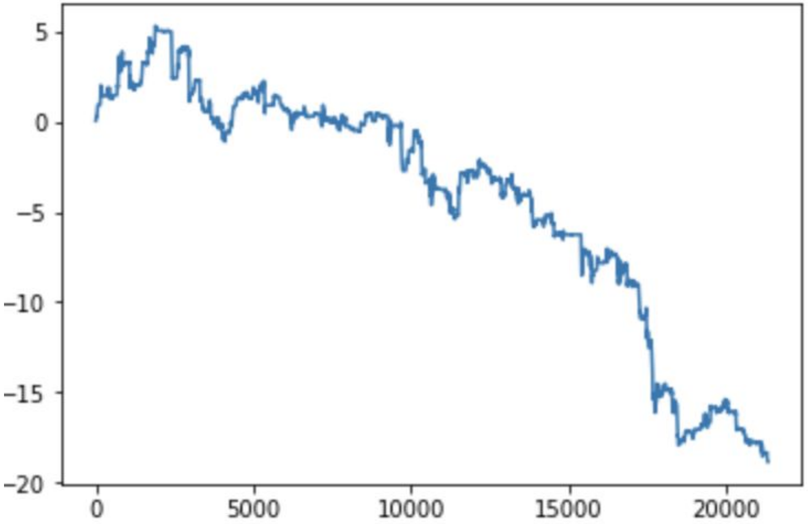
Current Stock Price: $34.01

TAKEAWAYS:

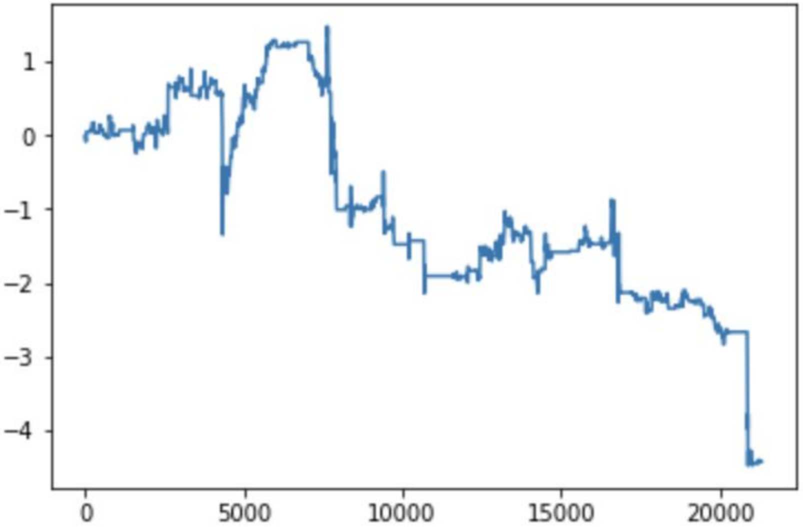
- High Fluctuation from results/Intraday Market Risk
- Lower predictive capabilities
- Overfitting and Multicollinearity
- Different Sector than previous example

Results: Profit/Loss 3 Days (10:00am-12:00pm)

VRSK: $-18.865

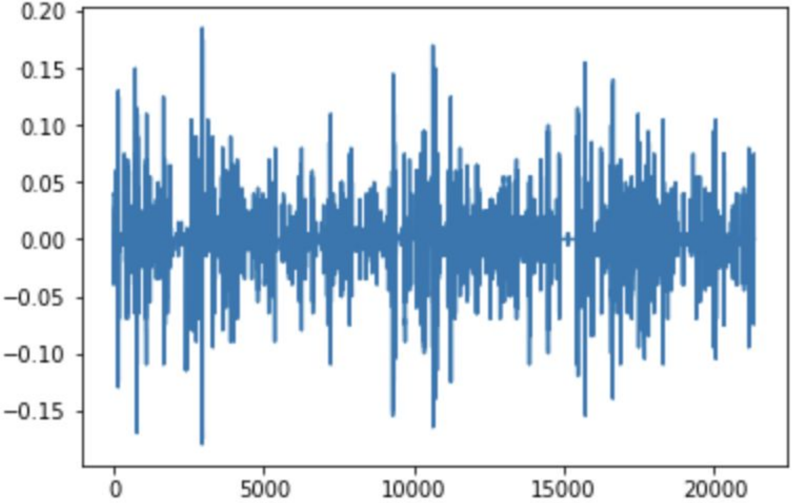


VIA: -4.43

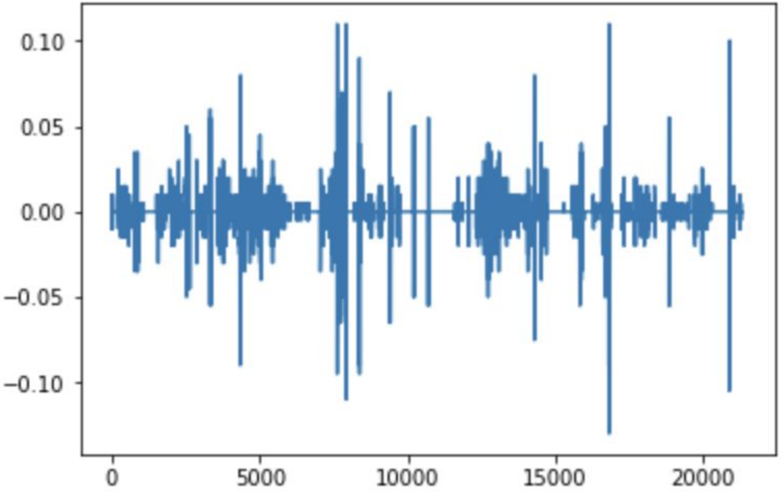


Results: Returns 3 Days

VRSK

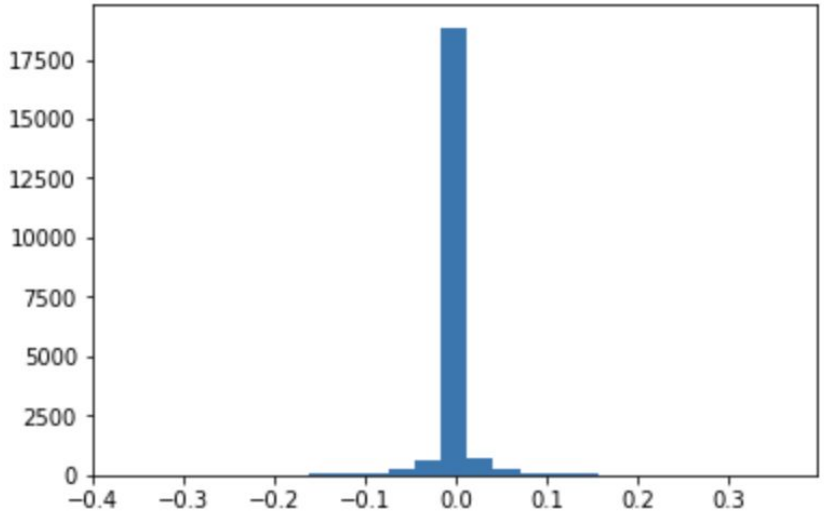


VIA

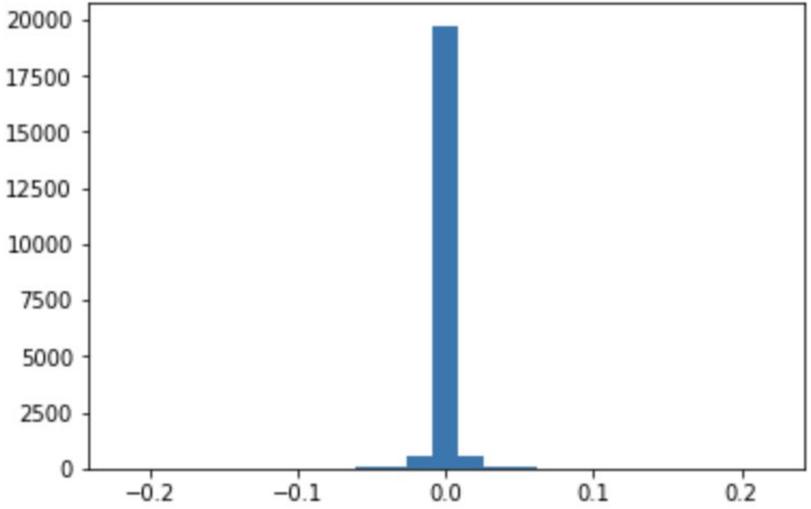


Results: Histogram of Returns 3 Days

VRSK



VIA



Results: Across tested Pairs (1/29-1/31) *10am-12pm

Pair of Stocks from NASDAQ 100	Profit (In $)
BKNG, NWSA	82.785
NTES, NWS	-26.605
NWS, PDD	1.34
NWS, VIA	2.96
PDD, VIA	2.395
HSIC, VIA	-2.595
IDXX, VIA	-1.97

Outline

- Review since midterm
- Model selection
- Simulation
- Results
- Discussion and future work

Discussion

Strengths:

- Made money even using a rudimentary model
- Able to capture returns from highly correlated stocks

Weaknesses:

- Lost money for some of our pairs
- Quantity is unspecified
- Bands are fixed

Future work

- Determine optimal quantities to trade on
- Optimal band selection
- Pull more data to train on
- Update models and pairings (at minimum) biannually
- Build our own simulator

References

[1] Cartea Alvaro, Jaimungal Sebastian, Penalva José (2015). Algorithmic And High-Frequency Trading.

[2] Almgren Robert, Chriss Neil(1999). Optimal Execution of Portfolio Transactions.

[3] Elliott, Robert & van der Hoek, John & P. Malcolm, William. (2005). Pairs Trading. Quantitative Finance.

Final pairs traded: (BKNG, NWSA), (FOX, NWS), (HSIC, VIA), (IDXX, NWS), (IDXX, VIA), (MELI, NWS), (NTES, NWS), (NTES, VIA), (NWS, PDD), (NWS, ULTA), (NWS, VIA), (NWS, VRSK), (PDD, VIA), (VIA, VRSK)

Thank you!

Questions?