Statistical Arbitrage in Small- Cap U.S. Stocks

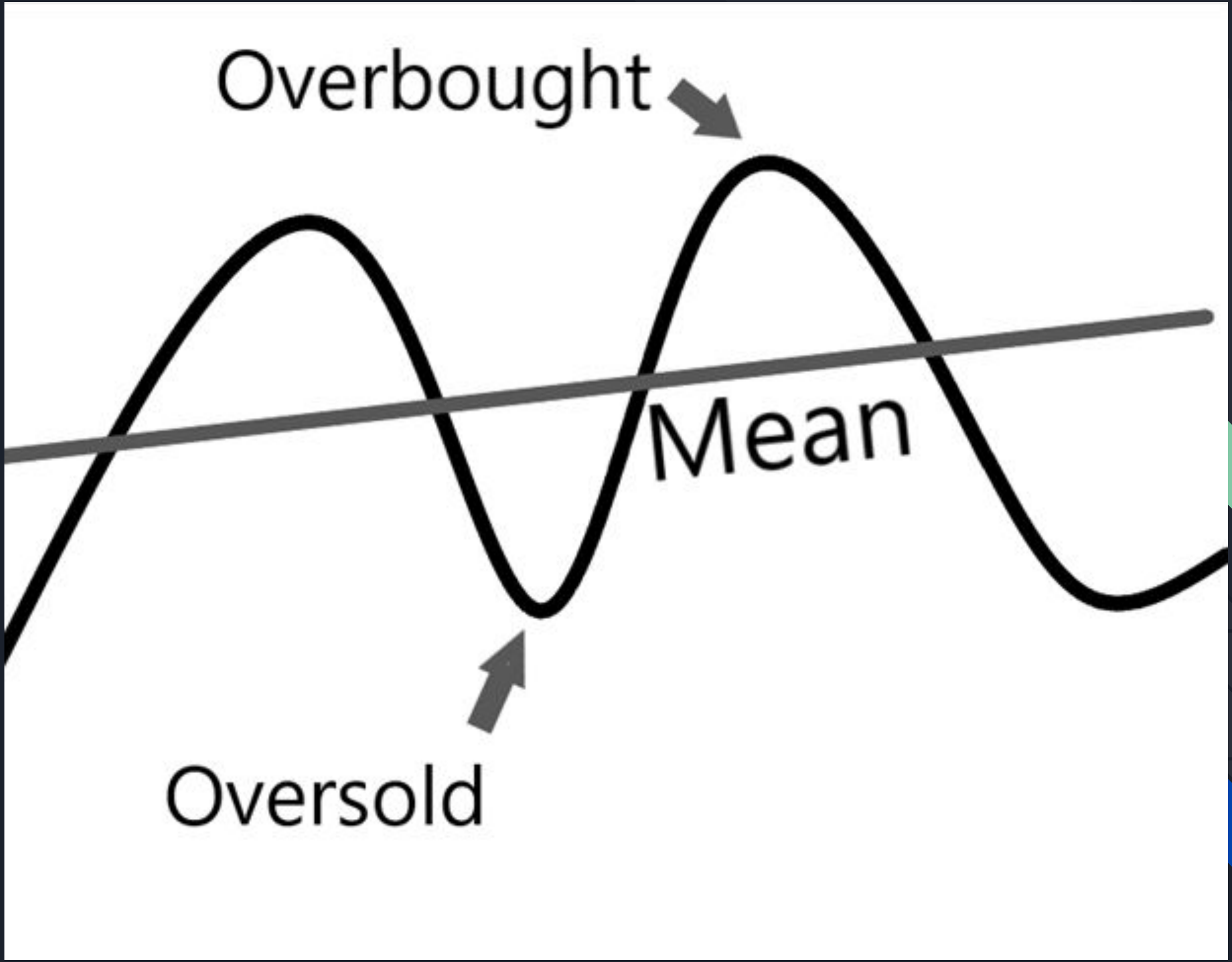
Brian Seabrooks and Trishiet Ray

Approach and Intuition behind Strategy

- Statistical Arbitrage strategies rely on the principle of mean-reversion, the strategies "bet" that mean reversion will occur within a short term time horizon. We considered this phenomenon within the basket of Russell 2000 stocks

- Implementing the strategies involve first using some sort of factor model to model the returns of a basket of stocks, allowing stock returns to be dissected into *systemic* and *idiosyncratic* components

- The way the factor model is designed is so that the expected value for the idiosyncratic component is 0


Overbought
Mean
Oversold

Dataset

- Russell 2000 US equities data from 2018 - 2020 from The Center for Research in Security Prices Dataset via Wharton Research Data Services
    - Wanted to include small cap stocks (vs. S&P 500)

- Data Availability Issues - 72 stocks had missing price data for these periods and were excluded from analysis.
    - They did not go under but were simply missing data

Standardized Daily Returns Matrix

	JJSF	ELA	PLXS	HNGR	AMRC
[1,]	-0.089120308	0.60311630	0.21772236	-0.46811642	0.3426483
[2,]	-0.654889610	0.47100865	-0.97527510	-0.26920381	-0.5897163
[3,]	-0.216518503	0.32978836	0.30929685	-0.32965631	1.0553831
[4,]	-0.043646011	-0.07529936	0.07041025	-0.04034122	-1.4824152
[5,]	0.528925857	1.55118702	0.63192236	0.61978084	-0.9644249
[6,]	-0.151105672	0.71233052	-0.41895882	-1.81425270	0.9805641
[7,]	-0.287249187	-1.69887751	-1.86839853	-1.39755887	0.4710536
[8,]	-0.040213053	1.91301018	-0.56714195	-1.30228586	0.3160674
[9,]	0.003925877	1.50132359	0.73754063	0.80664029	1.2298637
[10,]	-0.029493311	-0.36487246	-0.59794254	-0.16435566	-0.7180047
[11,]	0.245602883	0.36076054	1.65796096	0.89905982	1.7422555
[12,]	-0.185383685	-0.24717082	0.48399983	-0.71590673	0.3851682
[13,]	-0.065360916	-0.48428404	-0.18505477	-1.46480258	0.6348914
[14,]	0.068628601	0.17603360	0.04279031	0.27471055	-0.0430801
[15,]	0.180941192	-0.65839115	-0.99432676	-0.36213686	-1.8876303

$$Y_{ik} = \frac{R_{ik} - \overline{R}_i}{\overline{\sigma}_i}$$

$$\overline{R}_i = \frac{1}{M} \sum_{k=1}^{M} R_{ik}$$

$$\overline{\sigma}_i^2 = \frac{1}{M-1} \sum_{k=1}^{M} (R_{ik} - \overline{R}_i)^2$$

Choosing Our Factor Model

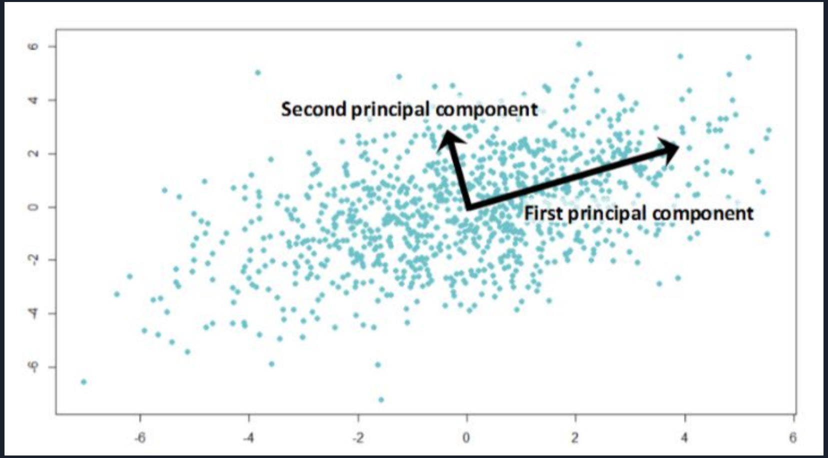
- There are many choices of factor models but they all reflect the same basic idea

$$\frac{dP_t}{P_t} = \alpha dt + \sum_{j=1}^{n} \beta_j F_t^{(j)} + dX_t,$$

- We chose to use principal components analysis to find the factors

Choosing Our Factor Model, continued....

- Principal Components seeks to identify the “components” of highest variance within a data set
- Here our “samples” are the dates under consideration, and the “variables” are the daily returns for each stock within our universe



Choosing Our Factor Model, continued....

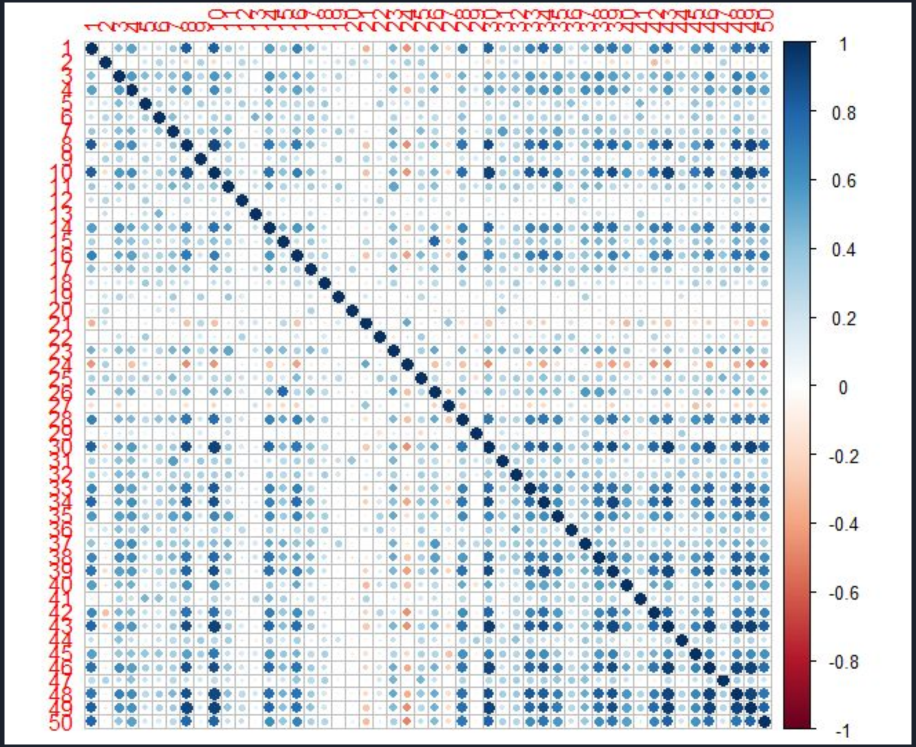
- After Identifying the Principal Components, we then construct hypothetical portfolios consisting of specific combinations of each of the stocks in our universe for each of the principal components that we'll use

$$Q_i^{(j)} = \frac{v_i^{(j)}}{\overline{\sigma}_i}.$$

$$F_{jk} = \sum_{i=1}^{N} \frac{v_i^{(j)}}{\overline{\sigma}_i} R_{ik} \quad j = 1, 2, ..., m.$$

- Using PCA to create the factors has two main advantages: the first is that it allows us to arrive at a set of *uncorrelated* factors, and the second is that it doesn't require us to make assumptions about the factors that drive stock returns, allowing us to design our factors more empirically
- However, one of the challenges with using PCA to design the factors is that you must select how many of the factors to use in the final model

Correlation Matrix

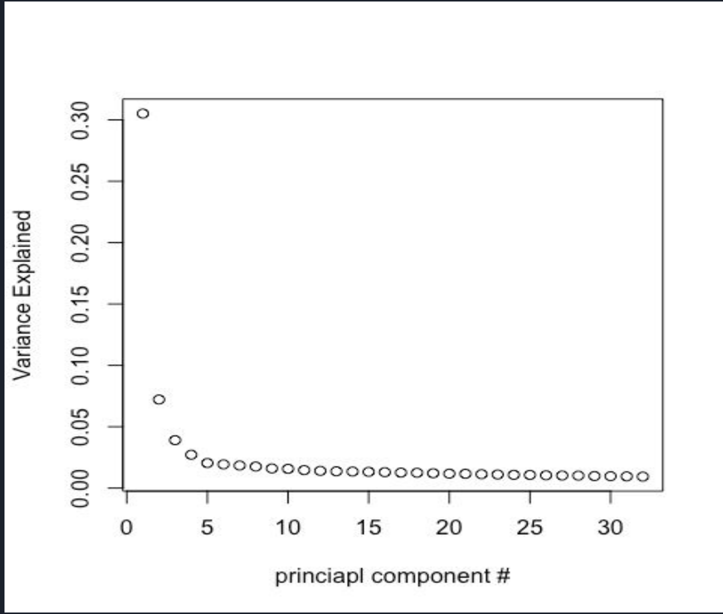


- Our correlation matrix shows the correlations of daily returns for each ticker. This is what is what we run our PCA on!

Selecting the number of factors to use

- **There are several trade offs when deciding how many factors to include**

More Components:	**Less Components:**
Factor Model describes data better	Idiosyncratic Component appears smaller
Controls Risks Better	Trading and Hedging are harder
Lowers Bias, but increases Variance of model predictions	Lowers Variance, increases Bias of model predictions

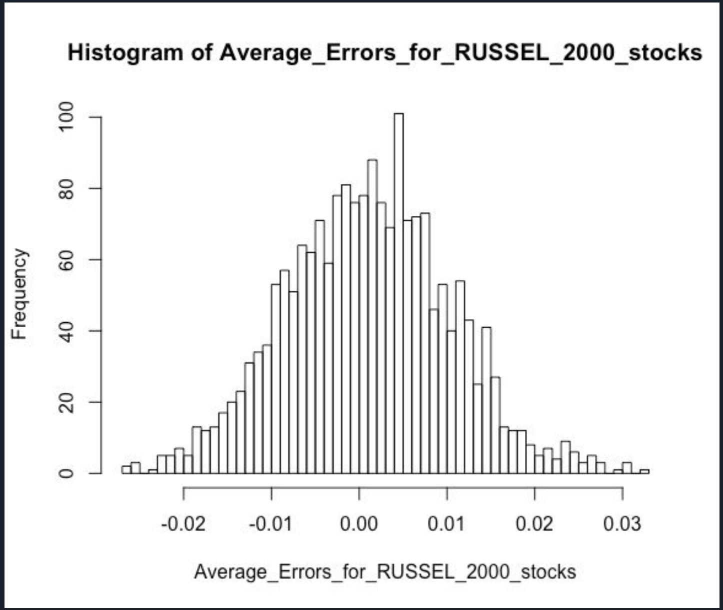


Selecting the factors, and then refining the basket of stocks

- We chose to set a threshold for variance explained (between 65% and 85%), and select the amount of principal components to use within out model
- This allows us to model the stock returns fairly robustly, while still controlling against overfitting, and also leaving room to model the idiosyncratic component
- Once we have the factors, we can refine the basket of stocks to include stocks whose returns are best described by the model but still have substantial idiosyncratic components

Refining the Basket of Stocks

- We want the real returns for a stock to be *reasonably* close to the factor model's prediction, but *not too* close to the prediction, because we want to model mean reversion
- We pick N stocks clustered close to the center of the histogram, but not *too* close to the center
- These stocks are well described by the factor model but still have strong idiosyncratic behavior



Modeling the Idiosyncratic Component of the Returns for the stocks we've selected

- We have now modeled the *systemic component* of the returns for these N stocks we've selected, we must now model the *idiosyncratic component*. Because we only bet on mean reversion over short periods of time, we assume that the drift is 0

$$\frac{dP_t}{P_t} = \alpha dt + \sum_{j=1}^{n} \beta_j F_t^{(j)} + dX_t,$$

↓

$$\alpha_i dt + dX_i(t).$$

Modeling the Mean Reverting Process

$$dX_i(t) = \kappa_i \left(m_i - X_i(t)\right) dt + \sigma_i dW_i(t), \quad \kappa_i > 0.$$

$$E\left\{dX_i(t) | X_i(s), s \leq t\right\} = \kappa_i \left(m_i - X_i(t)\right) dt \ .$$

$$E\left\{X_i(t)\right\} = m_i \ \text{ and } \ Var\left\{X_i(t)\right\} = \frac{\sigma_i^2}{2\kappa_i}.$$

- Because of how we fit the factor model, the residuals of the factor model are all forced to be 0. This is clearer when looking at the discrete version of the Orienstien Uhlembeck Process:

$$X_k = \sum_{j=1}^{k} \epsilon_j \quad k = 1, 2, ..., 60,$$

Modeling the Mean Reverting Processes, continued...

- The regression "forces" the residuals to have mean 0, so we have that $X_{60} = 0$
- Estimation of the OU parameters is done by solving an auto-regressive model

$$X_{n+1} = a + bX_n + \zeta_{n+1}, \quad n = 1, ..., 59.$$

- Finally we can find the parameters for our Orienstien-Uhlembeck processes:

$$\begin{aligned}
\kappa &= -log(b) * 252 \\
m &= \frac{a}{1-b} \\
\sigma &= \sqrt{\frac{\text{Variance}(\zeta) \cdot 2\kappa}{1-b^2}} \\
\sigma_{eq} &= \sqrt{\frac{\text{Variance}(\zeta)}{1-b^2}}
\end{aligned}$$

Generation of Trading Signals

Finally we can discuss how trading signals are generated

- On each day, all parameters are re-estimated using a 60 day look back period
  - We calculate the S-score for each stock
- The theoretical S-score is: $$s = \frac{X(t) - m}{\sigma_{eq}}$$
- But since $X(t) = X_{60} = 0$,

$$s = \frac{-m}{\sigma_{eq}} = \frac{-a \cdot \sqrt{1 - b^2}}{(1 - b) \cdot \sqrt{\text{Variance}(\zeta)}}.$$

- We enter a long/short position in stock i if S_i is below/above a certain threshold, we choose these thresholds
- We must complete the position by buying/selling other stocks so that the portfolio Betas remains at 0

Setting the S-score cutoffs for trading:

- We enter a long/short position in stock i if S_i is below/above a certain threshold, we choose these thresholds
  - Let's say the cutoff to enter a long position is if S_i is below B, and the cutoff to enter a short position is if S_i is above C

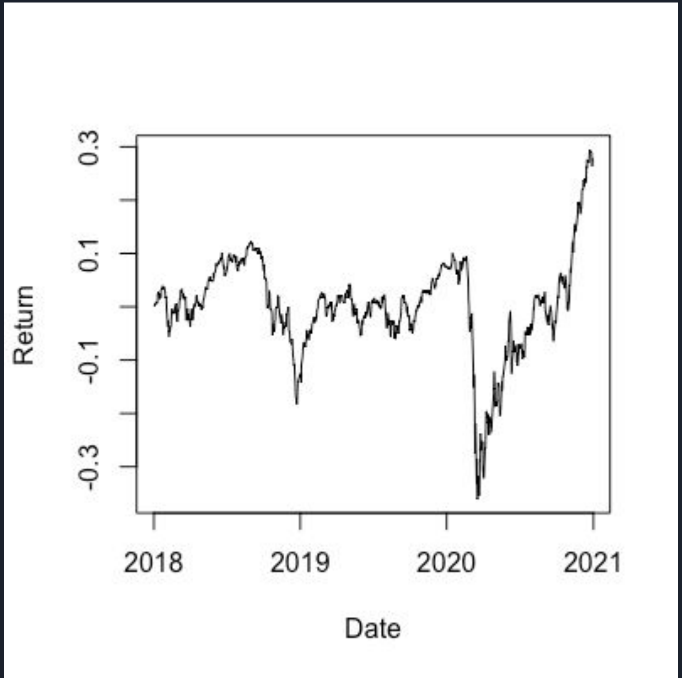
- As you decrease B, or increase C, the model will generate less trading signals because it's *harder* to meet the criteria for entering a trade. This will reduce the amount of trades that are entered, but in theory it should mean that you're more confident in the trades that you do enter. However because you're entering less trades, you benefit less from averaging.

It’s important to remember what parameter choices we make, and also recognize that we can tune these

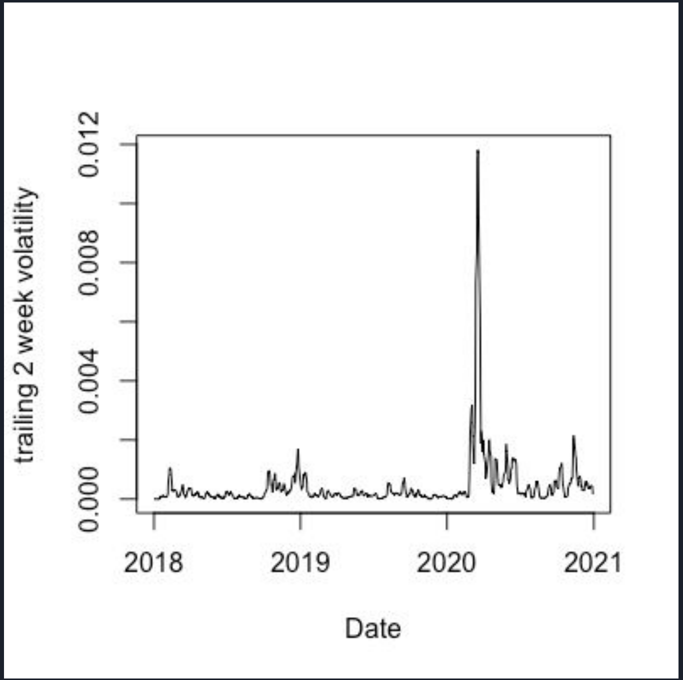
Look Back Period, M	We have M set at 60 days, but many choices can perform well
Threshold of Variance Explained by PCs	We consider thresholds between 65% and 85%
What percent of the total stocks in the universe to use for trading, and how to select them	We use the 25% of stocks in the universe that have average prediction errors that are *moderately* close to 0
The lag with which you model the OU process	We just choose a lag of 1, this is the simplest
Cutoff for entering long position, C	We have this at 1.25
Cutoff for entering short position, U	We have this at 1.25

Analyzing Results, looking closer at the Russell 2000

RUSSELL 2000 Returns 2018 - 2020

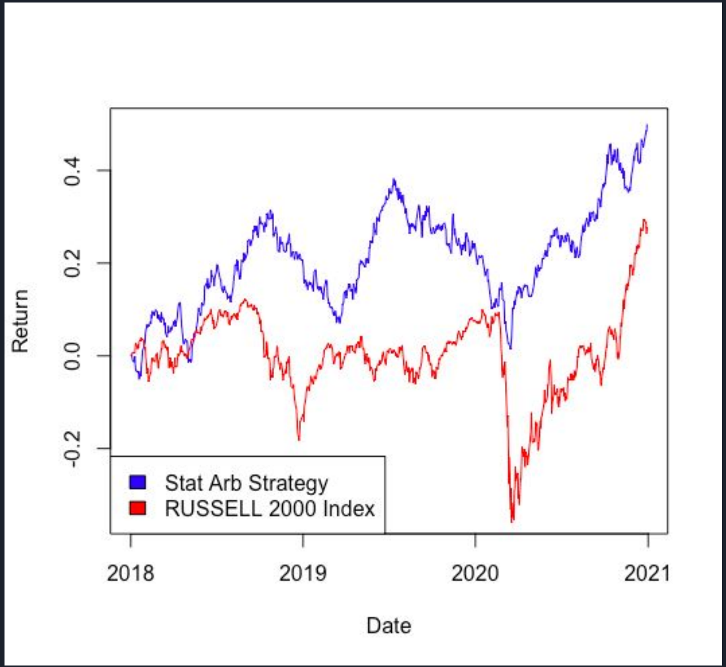


RUSSELL 2000 Trailing Volatility



Analyzing Results

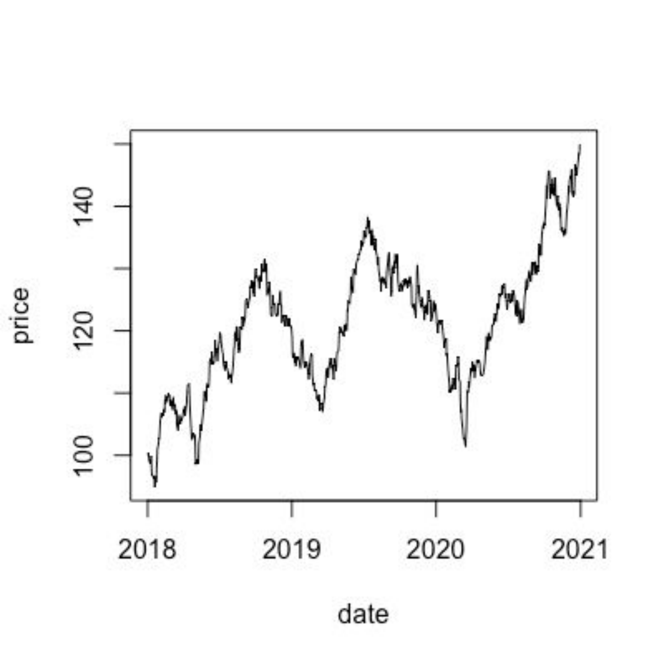
Stat Arb Returns vs. RUSSELL 2000



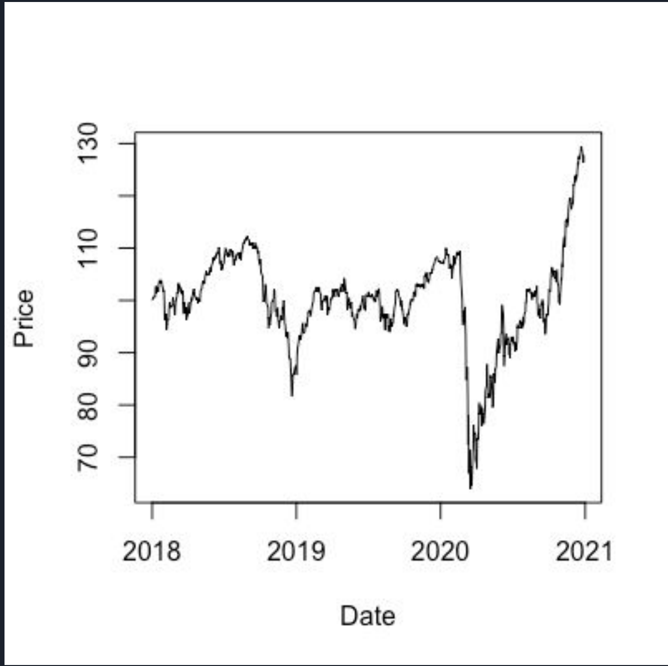
- Stat Arb achieves a Sharpe Ratio of 1.238, compared to the Index's .65
- Stat Arb returns seem more or less uncorrelated with index
- "Max drawdown" for strategy is much less that that for index
- We can see that the Stat Arb strategy outperforms the index the most when volatility is high

Visualizing and Comparing Profit and Loss and Maximum Drawdown for Stat Arb vs. index

Stat Arb Strategy



RUSSELL 2000



Risk Considerations, Implementation Difficulties

- Crowding into Statistical Arbitrage Based portfolios, causing hidden risks
  - Through focusing on small caps rather than large caps (where many stat arb strategies are applied) we think we mitigate this risk by a bit
- Lots of trading can be involved in hedging the positions properly
  - This can incur trading costs, and errors that aren’t accounted for in the model
  - Many of the assets that the strategy uses for trading/hedging are not purchasable in the quantities that we may like to purchase them
  - This can make it much harder to effectively control the risks of the portfolio
- During periods of very high market volatility correlations across assets spike significantly, meaning that the hedging done by our strategy is less effective
- Dealing with bugs in code (like having an initial row of all 0s in returns matrix)

Extensions and Enhancements

- The parameters we discuss earlier can be continually tuned
- Different Measures of volatility can be considered by the model
    - Volatility measures that consider “fat tails” and “volatility clustering”
    - This could help to more effectively control the portfolio’s risk
- The strategy could incorporate assets outside the RUSSELL 2000 index to hedge portfolio risks
- Alternative factor models can be explored
- Signal generation could consider not just mean reversion for a given stock, but also other indicators that might be relevant to it it’s short term idiosyncratic returns

Questions?

Appendix: Factor Based Decomposition of stock returns

Given that we have:

$$\overline{\beta}_j = \sum_{i=1}^{N} \beta_{ij} Q_i = 0, \quad j = 1, 2, ..., m.$$

We can show that:

$$\begin{aligned}
\sum_{i=1}^{N} Q_i R_i &= \sum_{i=1}^{N} Q_i \left[\sum_{j=1}^{m} \beta_{ij} F_j \right] + \sum_{i=1}^{N} Q_i \tilde{R}_i \\
&= \sum_{j=1}^{m} \left[\sum_{i=1}^{N} \beta_{ij} Q_i \right] F_j + \sum_{i=1}^{N} Q_i \tilde{R}_i \\
&= \sum_{i=1}^{N} Q_i \tilde{R}_i
\end{aligned}$$

This is exactly what our goal is! To only be exposed to the idiosyncratic component of stock returns.