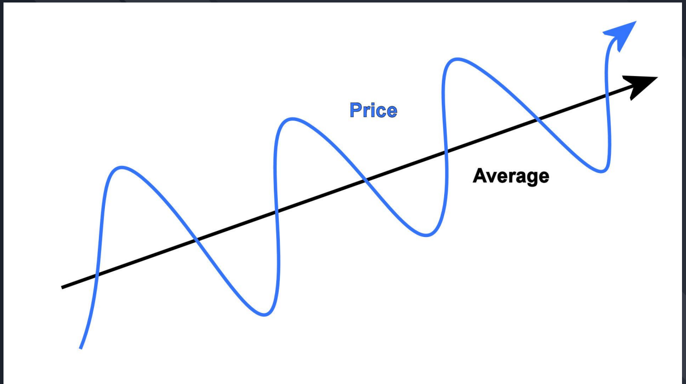
Statistical Arbitrage in U.S. equities

By Brian Seabrooks and Trishiet Ray

Overview of Statistical Arbitrage Strategies

- A formalization of general “pairs trading”, in which trader takes a long position in a stock that he believes is underpriced relative to another stock. Takes an equivalent short position in the corresponding stock.
- This aims to create what’s called a 0-Beta strategy, Trading book has 0-Beta
- Seeks to dissect stock returns into systematic components and idiosyncratic components
    - Rather than try to predict the systematic components (this can almost be thought of as trying to predict where the market is going), the modeling is focused on the idiosyncratic component of stock returns
- Traders doing statistical arbitrage projects focus on modeling the idiosyncratic components of returns



Dataset

Wharton Research Data Services: CRSP

- The Center for Research in Security Prices, LLC (CRSP) maintains the most comprehensive collection of security price, return, and volume data for the NYSE, AMEX and NASDAQ stock markets.

- Universe: 10 buckets of 3 - 10 stocks each, manually constructed of intuitively “similar” stocks
- Timeframe: 1 - 7 days
- Data from 2015 to 2020

Dataset Format

```
PERMNO,date,NAMEENDT,SHRCD,EXCHCD,SICCD,NCUSIP,TICKER,COMNAM,SHRCLS,TSYMBOL,NAICS,PRIMEXCH,TRDSTAT,SECSTAT,PERMCO,ISSUNO,HEXCD,HSICCD,CUSIP,DCLRDT,DLAMT,DLPDT,DLSTCD,NEXTDT,PAYDT,RCRDDT,SHR
10026,20190102,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,139.12000,144.03999,141.00000,112825,-0.024829,140.75999
10026,20190103,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,140.27000,145.91000,143.02000,84331,0.014326,143.02000,1
10026,20190104,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,143.16479,147.36501,144.84000,88233,0.012725,144.84000,1
10026,20190107,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,142.83000,147.87000,145.41000,79539,0.003935,145.20000,1
10026,20190108,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,145.50000,149.30231,148.70000,70200,0.022626,148.28000,1
10026,20190109,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,146.09500,151.00000,147.71001,63591,-0.006658,147.16000,
10026,20190110,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,146.12000,150.00000,148.91000,66117,0.008124,148.91000,1
10026,20190111,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,147.71001,150.44000,148.89000,58050,-0.000134,148.62000,
10026,20190114,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,147.70000,149.56500,147.70000,92311,-0.007992,147.70000,
10026,20190115,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,145.20000,148.50000,146.00000,89393,-0.011510,145.61000,
10026,20190116,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,0,,,,,,,,,,,,,1,6,44,2,142.99001,148.19501,143.84000,105068,-0.014795,143
10026,20190117,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,0,,,,,,,,,,,,,1,6,43,2,144.00000,147.09000,145.75999,85383,0.013348,145.8
10026,20190118,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,143.61000,149.48000,147.24001,98586,0.010154,146.86000,1
10026,20190122,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,143.42000,147.70000,144.53999,112506,-0.018337,144.46001
10026,20190123,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,144.00000,146.72000,146.21001,138011,0.011554,146.06000,
10026,20190124,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,0,,,,,,,,,,,,,1,6,43,2,145.14999,146.91000,145.80000,90458,-0.002804,145.
10026,20190125,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,0,,,,,,,,,,,,,1,6,45,2,144.23000,147.44000,144.55000,90705,-0.008573,144.
10026,20190128,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,138.39999,143.13000,140.00000,156991,-0.031477,140.00000
10026,20190129,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,146.80000,162.84000,156.82001,251608,0.120143,156.82001,
10026,20190130,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,149.88000,159.60001,150.50000,87854,-0.040301,150.50000,
10026,20190131,,11,3,2052,46603210,JJSF,J & J SNACK FOODS CORP,,JJSF,311821,Q,A,R,7976,10433,3,2052,46603210,,,,,,,,,,,,,,,,,,,,,,,,,148.42999,154.38000,154.35001,109251,0.025581,154.28999,
```

Example - 2 Correlated Stocks

Modeling Approach

General Pairs Model

$$ln(P_t/P_{t_0}) = \alpha(t - t_0) + \beta ln(Q_t/Q_{t_0}) \ + \ X_t$$

Differential Form

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t,$$

Factor Decomposition of the Differential Form

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

Factor Based Decomposition of Stock Returns

- Our modeling approach relies on the notion that the returns for stocks can be “projected” onto a finite amount of factors
- After we do this projection, what we are left with is the idiosyncratic component of a stock’s returns: the drift and a stationary mean reverting process.

$$R_i = \sum_{j=1}^{m} \beta_{ij} F_j + \tilde{R}_i.$$

- A portfolio is “market neutral”, or neutral to the m risk factors under consideration if:

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

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.

Finally, how should we choose the factors?

Given that our goal is to find a set of M Factors, F_1,F_2, …., F_M such that if we project stock returns onto the factors we are left with only the idiosyncratic component of stock returns, selecting the factors wisely is obviously very important!

PCA approach to select risk factors

- Create a matrix, R, representing the N stocks in the “universe”, going back M + 1 days
- M represents the “look back period”. Many people take it to be ~60 days

$$R_{ik} = \frac{S_{i(t_0-(k-1)\Delta t)} - S_{i(t_0-k\Delta t)}}{S_{i(t_0-k\Delta t)}}, \; k = 1, ..., M, \; i = 1, ..., N,$$

- Now standardize the returns: $Y_{ik} = \frac{R_{ik} - \overline{R}_i}{\overline{\sigma}_i}$

PCA approach to select risk factors....

- Now that we have a matrix of standardized returns for the N stocks in our universe, going back M days, we can construct the empirical correlation matrix
- The empirical correlation between two stock returns is

$$\rho_{ij} = \frac{1}{M-1} \sum_{k=1}^{M} Y_{ik} Y_{jk},$$

- We now do principal component analysis on this empirical correlation matrix
- We'll end up with N eigenvectors with corresponding eigenvalues

PCA approach to select risk factors....

- After doing PCA on the empirical correlation matrix we have N eigenvectors and corresponding eigenvalues
- We can rank the eigenvectors by their corresponding eigenvalues

$$N \geq \lambda_1 > \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_N \geq 0.$$

- Each of the principal components correspond to a specific "portion of variance explained in the system"
- We denote the eigenvectors as

$$v^{(j)} = \left(v_1^{(j)}, ..., v_N^{(j)}\right) \ , \ \ j = 1, ..., N.$$

PCA approach to select risk factors

- Create the “eigen-portfolios” to arrive at risk factors

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

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

- These returns of the “eigen-portfolios” are uncorrelated with each other and will be the risk factors used in our model
- The returns of these eigen-portfolios will track the returns of a particular “risk factor”
- Now we are ready to return to the original problem of interest!

Stock Returns must be modeled relatively

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

$$\frac{dS_i(t)}{S_i(t)} = \alpha_i \, dt + \sum_{j=1}^{N} \beta_{ij} \frac{dI_j(t)}{I_j(t)} + dX_i(t),$$

- The idiosyncratic component of the stock return is: $\alpha_i dt + dX_i(t).$
- $alpha_i dt$ is the excess rate of return of the stock over the period in question
- $dX_i(t)$ is assumed to be the increment of a stationary stochastic process
    - This is the mean-reverting term, and it is the area of interest
    - It captures idiosyncratic fluctuations in a stock price
    - The way the parameters are fit, we must have $E[dX_i(t)] = 0$, which is exactly what we want

Mean Reversion Process

- All parameters in the model are re-estimated at the end of each trading day, to be used in the trading for the subsequent trading day
- The statistics for the residual/idiosyncratic process are continuously updated by assuming that the parameters are constant over the 60 day “look back” period
- A parametric method called the Orienstien Uhlemback process is used to model $dX_i(t)$, the parameters are easy to estimate

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

- This process is stationary and autoregressive with lag 1
- $dX_i(t)$ has unconditional mean 0, and conditional mean :

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

Mean Reversion Process - parameter estimation

- The parameters of the stochastic differential equation $dX_i(t)=$ …. , $alpha_i$, k_i, m_i, and $sigma_i$ are specific to each stock, and must be estimated
- We assume that these parameters vary slowly compared to the Brownian Motion Increment $dW_i(t)$ in the time window of interest
- If we assume that all parameters are momentarily constant we can write:

$$X_i(t_0+\Delta t) = e^{-\kappa_i \Delta t} X_i(t_0) + \left(1 - e^{-\kappa_i \Delta t}\right) m_i + \sigma_i \int_{t_0}^{t_0+\Delta t} e^{-\kappa_i (t_0+\Delta t - s)} dW_i(s)$$

- $X_i(t)$ is normal with:

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

Mean Reversion - Parameter Estimation

- K_i is called "the speed of mean reversion", $T_i = 1/K_i$ represents the characteristic scale for mean reversion
  - If K >> 1, then the stock reverts very quickly to its mean, and the effect of the drift term can be negligible
  - Stocks with quick mean reversion are our favorites!
- Now we do the regression: $R_n^S = \beta_0 + \beta R_n^I + \epsilon_n, \quad n = 1, 2, ..., 60.$
- Recall that:

$$\frac{dS_i(t)}{S_i(t)} = \alpha_i \, dt + \sum_{j=1}^{N} \beta_{ij} \frac{dI_j(t)}{I_j(t)} + dX_i(t),$$

- This tells us that: $\alpha = \beta_0/\Delta t = \beta_0 * 252.$

Mean Reversion - Parameter Estimation

- Now we can define the auxiliary process: $$X_k = \sum_{j=1}^{k} \epsilon_j \quad k = 1, 2, ..., 60,$$
- This can be viewed as a discrete version of X(t), the OU process that we are estimating
- 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 a lag 1 regression model

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

- Finally,

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

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

Next Steps and Anticipated Results

Next Steps:

- Decide on a specific universe of stocks
- Write code to continuously fit all parameter estimates for the stocks in our universe
- Decide on rules for the trading signals(these are tuning parameters)
- Begin trading
- Evaluate model performance

Anticipated Results:

- Returns for statistical arbitrage strategies tend to be modest, given that they seek to minimize exposure to all risk factors, and track idiosyncratic returns of stocks

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

- Similar strategies have achieved sharpe ratios of .9 to 1.44