Statistical Arbitrage on US ETFs

MS&E 448 Final Presentation

Shane Barratt Russell Clarida Mert Esencan Francesco Insulla
Cole Kiersznowski Andrew Perry[1]

Stanford University

May 5, 2020

[1] Authors listed in alphabetical order

Overview

Background

- Statistical arbitrage = short-term trading strategy that bets on mean-reversion of asset baskets (more later)
- The intuition of statistical arbitrage is based on the idea that the difference between what an equities' price is and what it should be is driven by idiosyncratic shocks
- Statistical arbitrage requires 3 steps:
  1. Finding asset baskets
  2. Prediction based on mean-reversion
  3. Portfolio construction

Market–Neutral Investments

- n assets, with prices $p_t \in \mathbb{R}^n_+$ at time period $t = 1, \ldots, T$
- Assume assets are hedged w.r.t. market, *i.e.*, each asset is actually 1 unit of the asset and $-\beta$ units of the market, where β is the correlation of the market returns with the asset returns
- Observation: **Investing (long or short) in any of these assets is market-neutral**

TAQ data

- 21 ETFs: SPY, MDY, DIA, XLK, XLV, XLF, XLP, XLY, XLU, XLE, XLI, XLB, QQQ, IVV, IWB, IWF, IJH, IJR, IWN, IWD, IVW, IVE
- Minute-level price data from NYSE TAQ WRDS for 2003-2020
- 1 million+ price points
- Used Yahoo Finance API to get corporate actions and adjusted for splits (ETFs do split!)

TAQ data

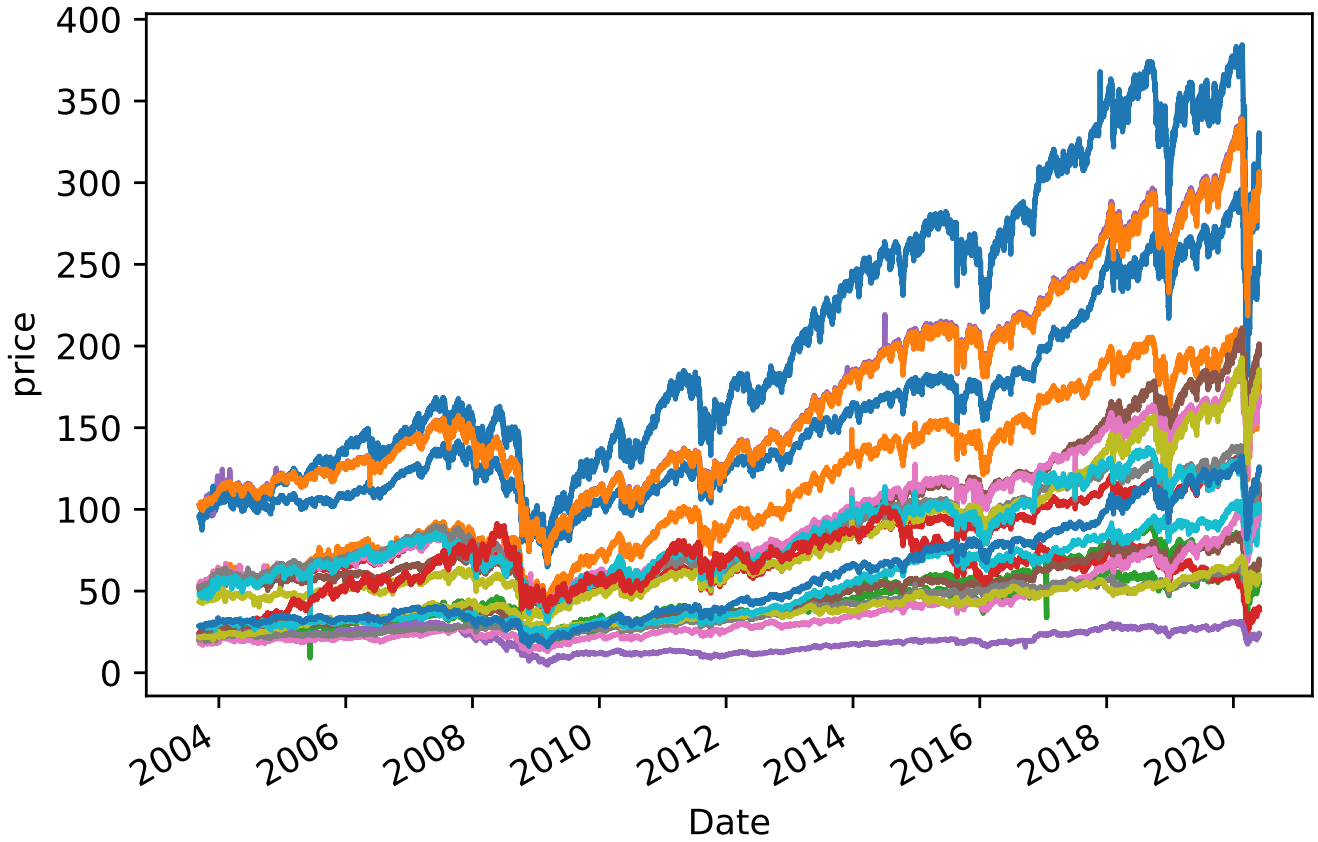

Figure: Index Prices Over Time

Hedging each ETF

- Each day, construct window of last 5 days
- Compute correlation of each ETF's returns with respect to SPY in window
- Construct market-neutral basket for each ETF

Finding Pairs

- Create asset basket for each of the 190 pairs via regression on window

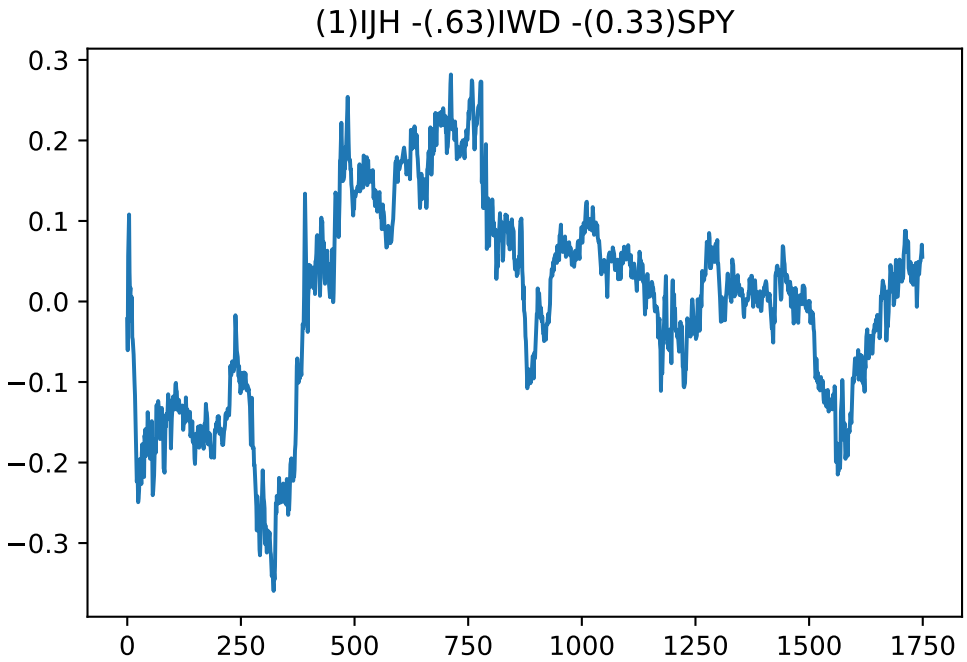

Figure: Example basket.

- Hope that some of the 190 asset baskets are mean-reverting

Modeling Residual Returns

- The residual returns are modeled with a **Ornstein–Uhlenbeck (OU)** process

- The OU process is mean-reverting which satisfies the assumption that prices return to basket means:

$$dX_t = \rho(\mu - X_t)dt + \sigma dB_t, \quad \rho, \sigma > 0 \tag{1}$$

  where ρ is the speed of mean reversion, μ is the long run average, σ is the instantaneous volatility, and B_t is a standard Brownian motion

- We allow for the assumption that over a short trading period that ρ, μ, and σ stay constant

AR(1) Process

- Constants ρ, μ, and σ in the OU model are calculated with a AR(1) process (Auto-Regressive with lag one).
- An AR(1) process is given by

$$X_{t+1} = \lambda + \phi X_t + \epsilon_t \tag{2}$$

- The interpretation of λ, ϕ, and ϵ as it relates to the OU process is

$$\begin{aligned} \lambda &= \mu(1 - e^{-\rho\delta t}) \\ \phi &= e^{-\rho\delta t} \\ \epsilon_t &\sim \mathcal{N}\left(0, \frac{\sigma^2}{2\rho}(1 - e^{-2\rho\delta t})\right) \end{aligned} \tag{3}$$

- We fit an AR(1) to all possible baskets.

Trading Signal

- Now that we are able to model stocks as an OU process we need a dimensionless trading signal
- We will use the distance that X_t is from the mean μ. Giving us the signal

$$s_t = X_t - \mu \tag{4}$$

- We use a linear bet size on the strength of the signal to invest in each basket. This allows us to make returns as the signal returns to zero.

Trading

- After we find the weights of each basket we can back out the weights on each of the ETFs
- Go long low baskets
- Go short high baskets
- Leverage around 2-3
- No TCs (major limitation)
- Implemented in numpy w/ vectorized operations. Takes about 10 minutes for full backtest

Results

Sharpe 1.244

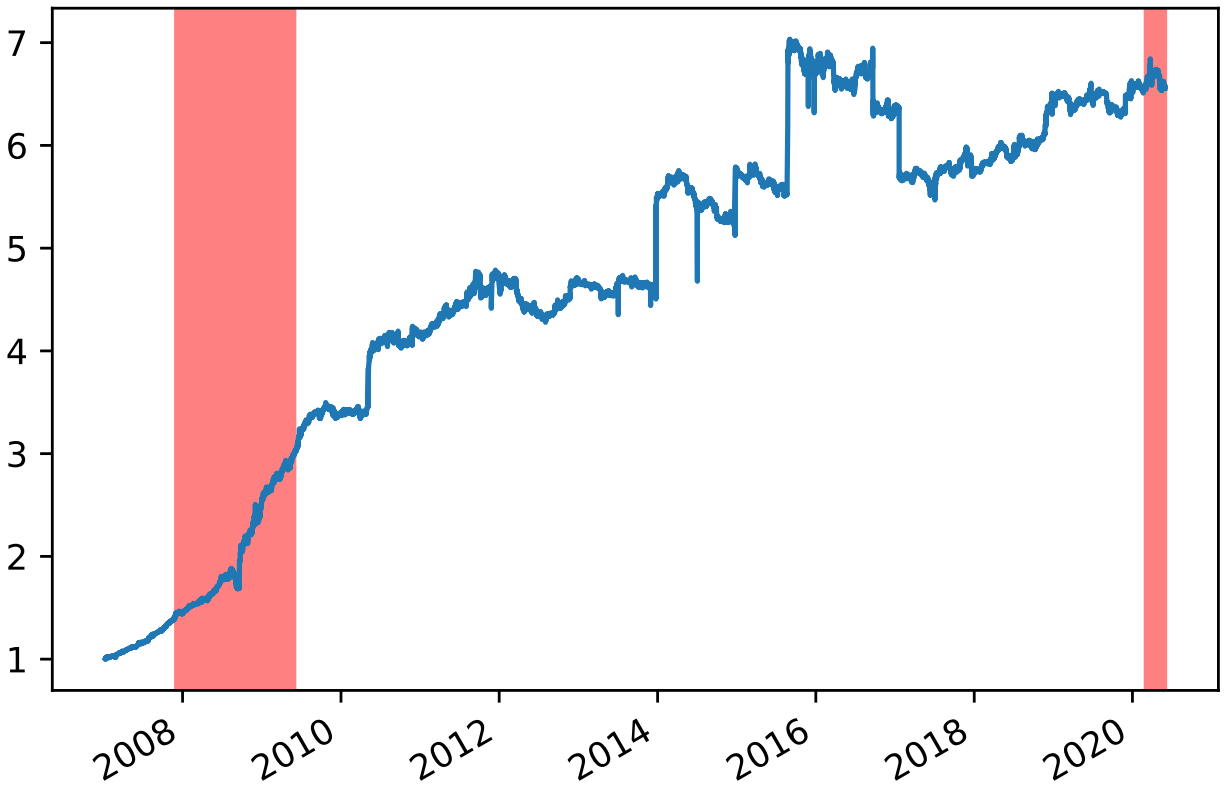

Figure: Daily Returns of Statistical Arbitrage Strategy with NBER Recession Bars

Drawdown

Max Drawdown 20%

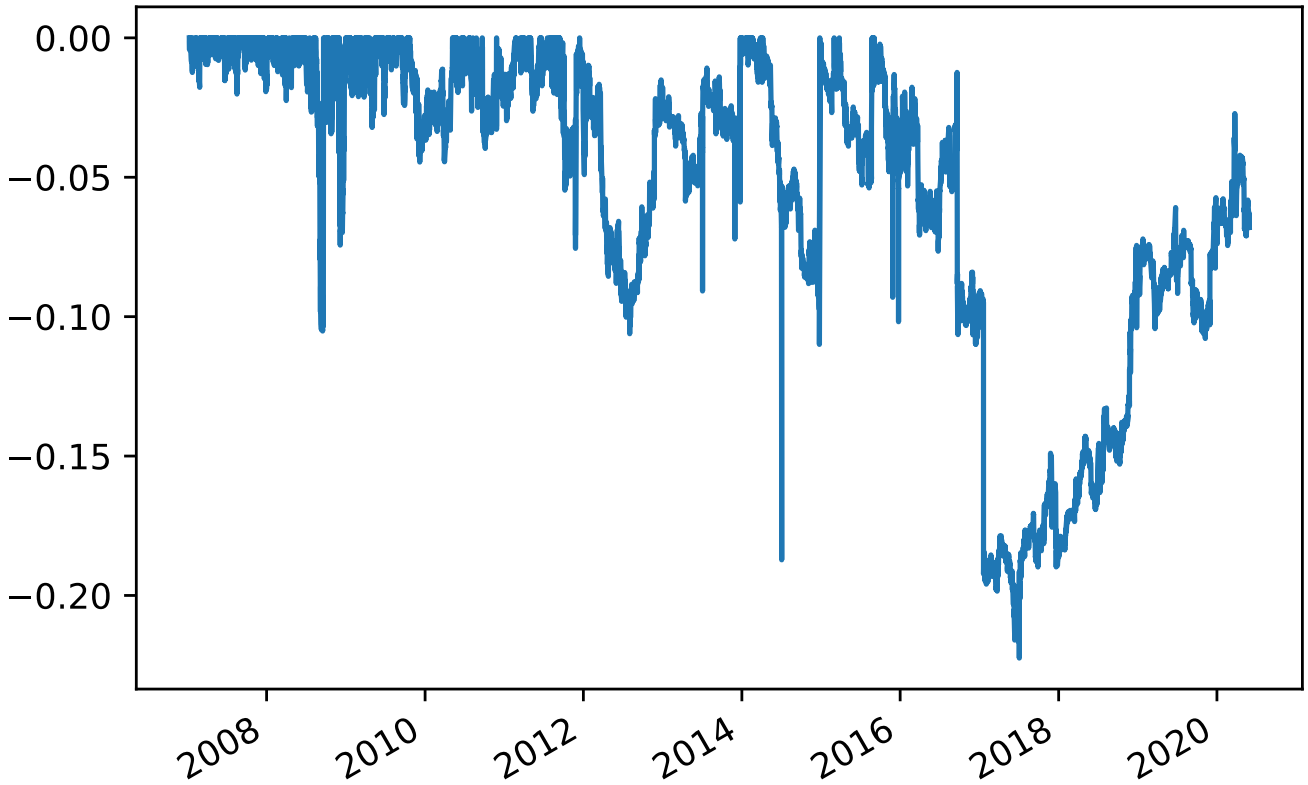

Figure: Drawdown

Daily Returns Regressed on the VIX

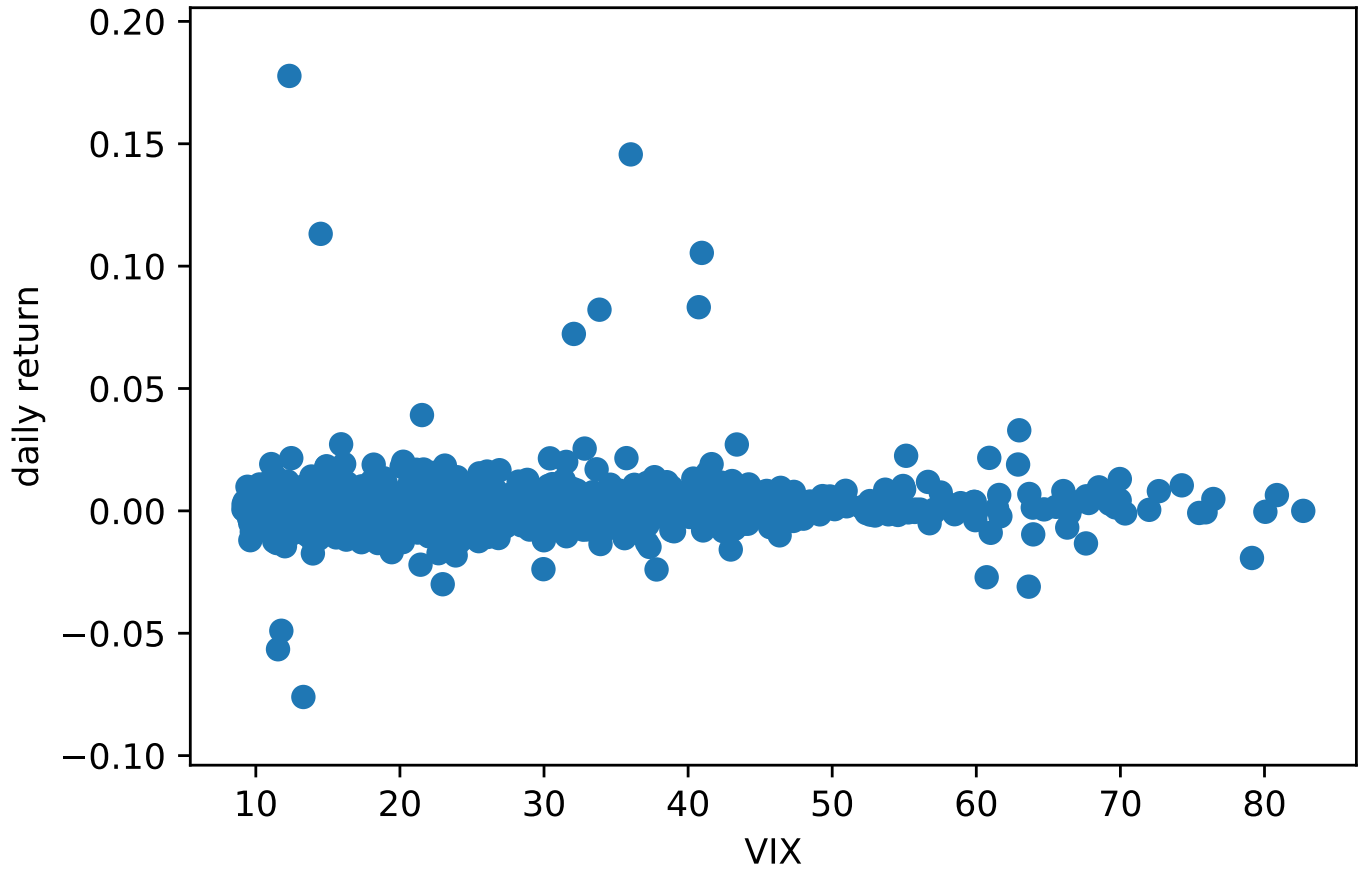

Figure: Daily Returns Regressed on the VIX

Daily Returns Regressed on the VVIX

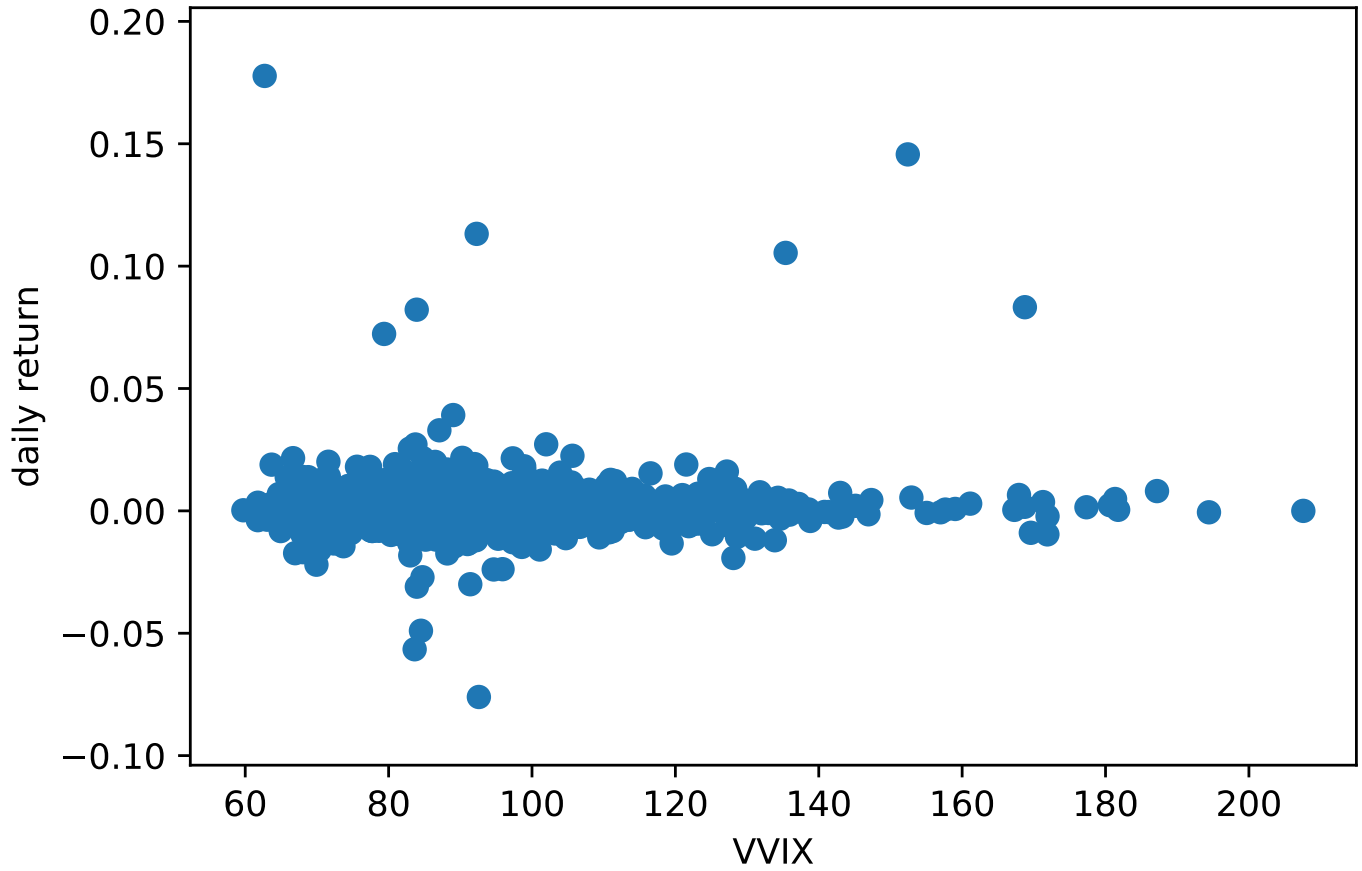

Figure: Daily Returns Regressed on the VVIX

Daily Returns Regressed on SPY

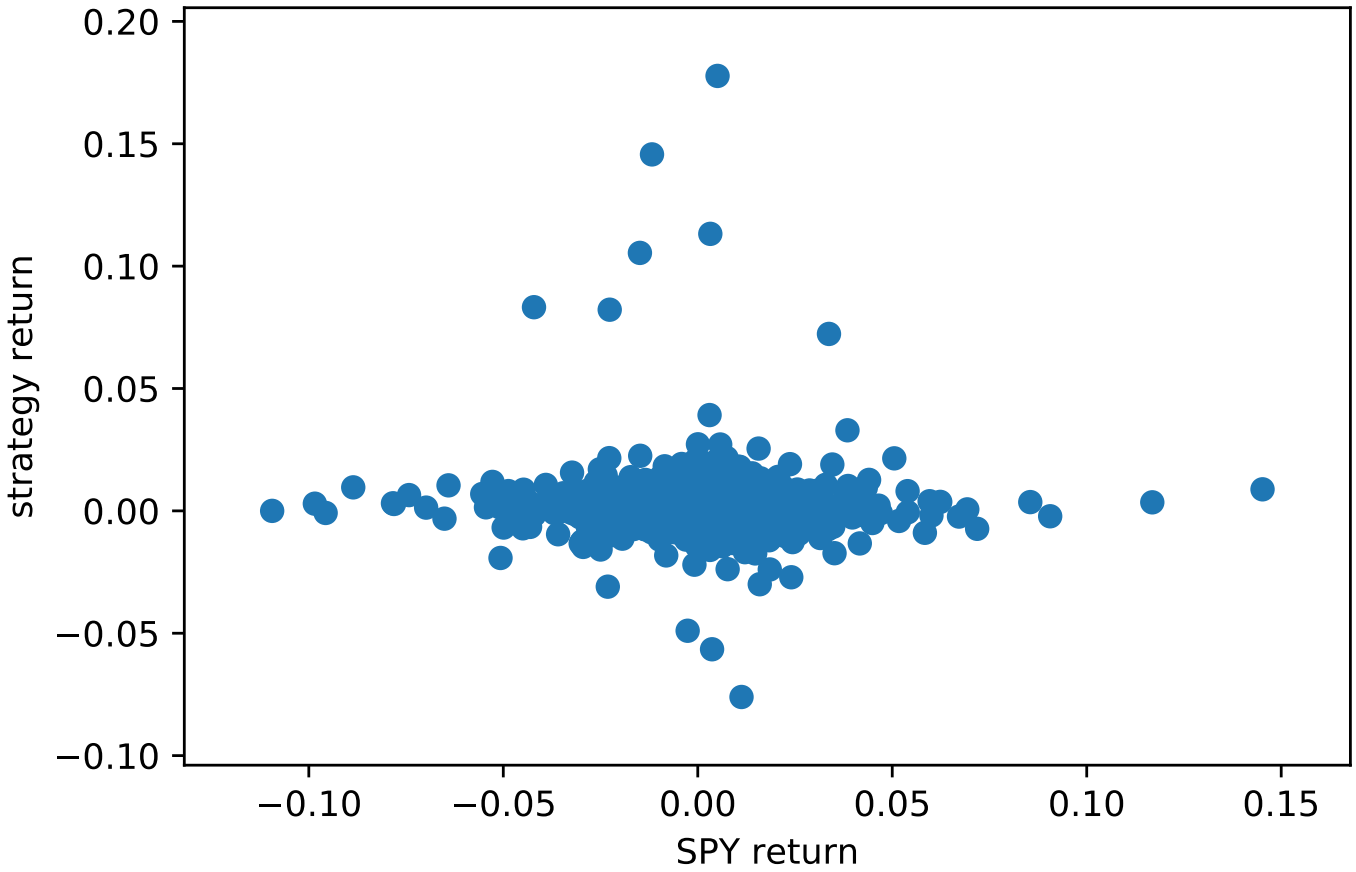

Figure: Daily Returns Regressed on SPY

Daily Returns on Shuffled Data

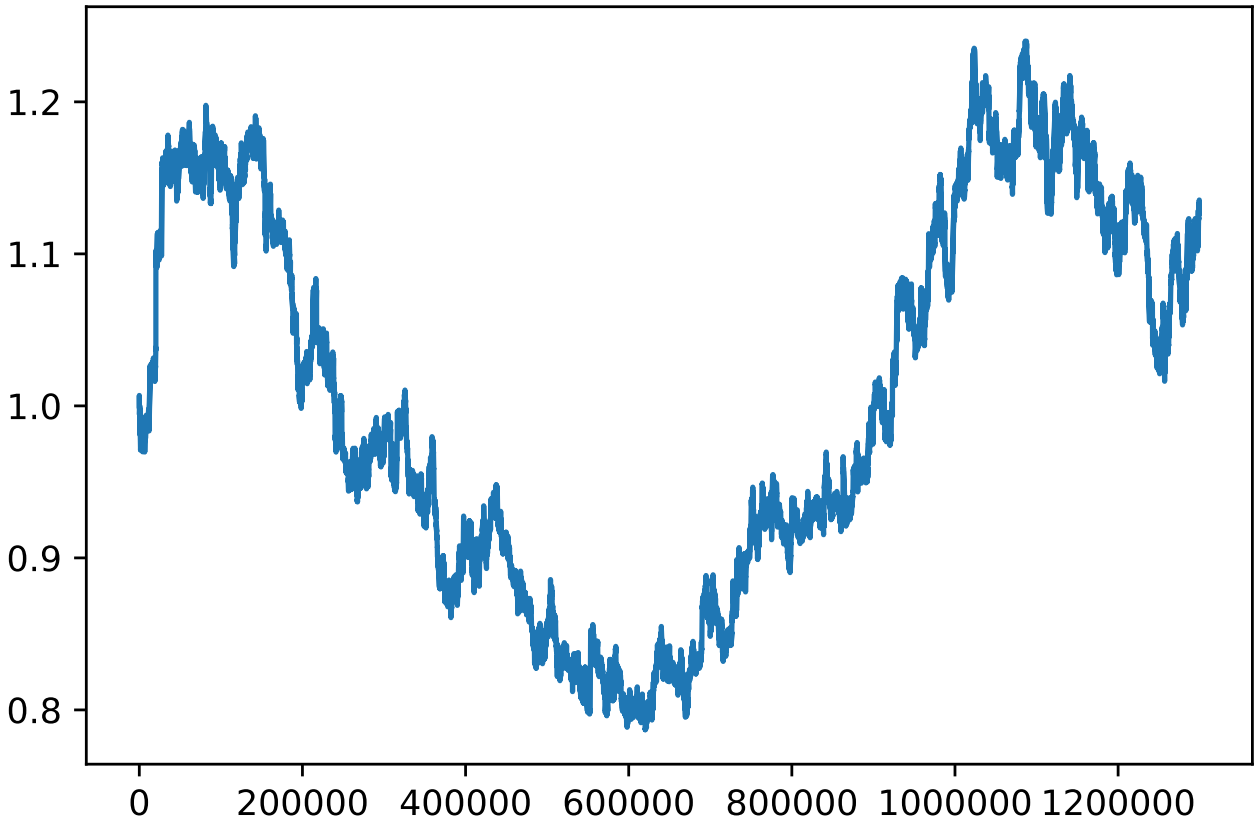

Figure: Daily Returns on Shuffled Data

Conclusion

Much of the struggle in this process is data acquisition and pipelines. While we believe the result to be compelling, we also know that structuring strong back-testing frameworks, acquiring and storing new data, and testing an array of signals is the minimum requirement for a stat arb strategy to even begin to hope to be competitive.

Questions?

Thank You

Appendix: A more complete approach

- Given a subset of returns, to find the optimal weights, we can maximize the "AR-ness", instead of using residuals or regression.

$$\text{Let } Z = X\alpha$$

$$\text{Assume } Z \sim \mathrm{AR}(1)$$

$$\text{Then } Z = \rho BZ + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_0^2)$$

$$\text{Regress } \Delta Z = \underbrace{\delta}_{\rho - 1} BZ + \varepsilon$$

$$\text{Get } \hat{\delta}, \sigma^2, \sigma_0^2, \mathrm{se}^2(\hat{\delta}), t_0 = \frac{\hat{\delta}}{se(\hat{\delta})}$$

$$\text{ADFuller: } H_0 \equiv \quad \delta = 0, \quad p \approx 2(1 - \Phi(t_0))$$

$$\boxed{\alpha^* = \underset{1^T\alpha = 1}{\operatorname{argmin}} - t_0}$$

Appendix: Comparison to naive approach

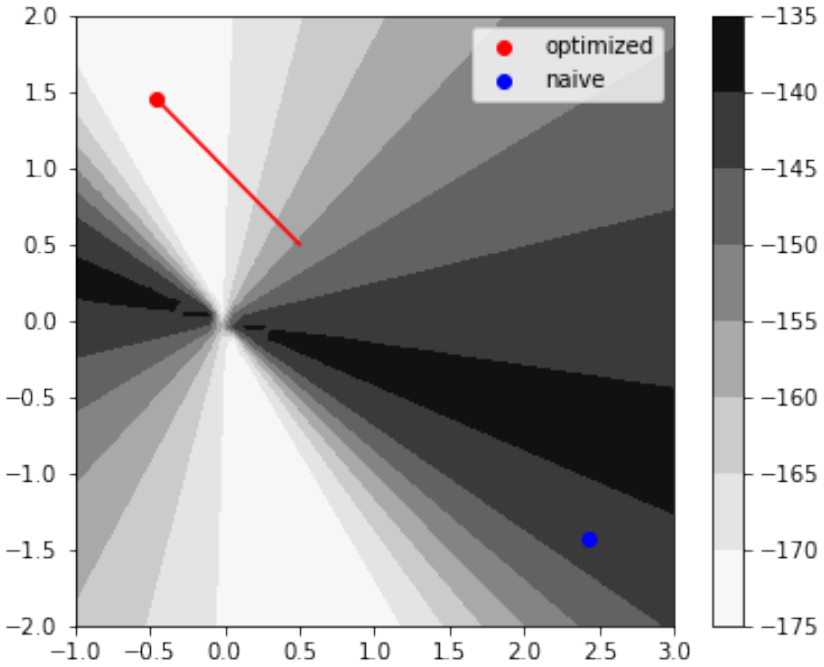

Figure: $-t_0$ contour plot

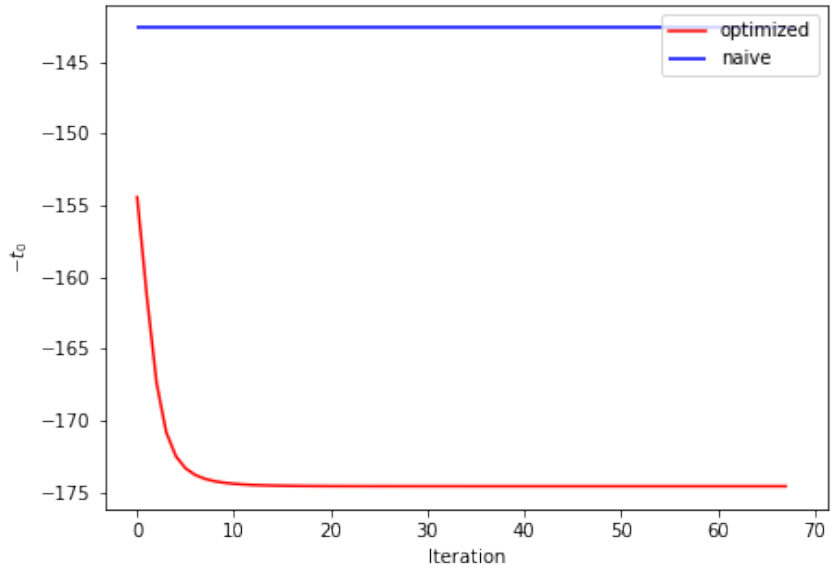

Figure: $-t_0$ vs iteration

Appendix: Across all pairs

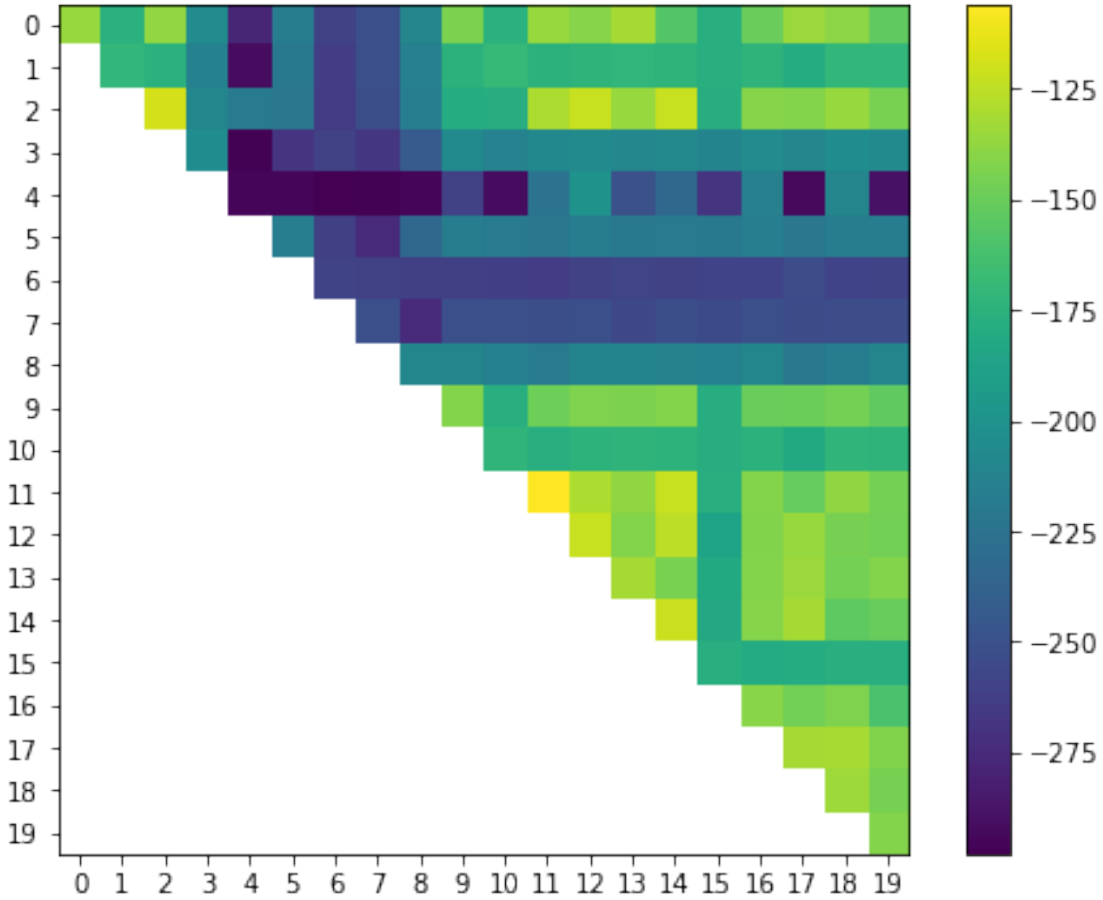

Figure: $-t_0$ for all pairs