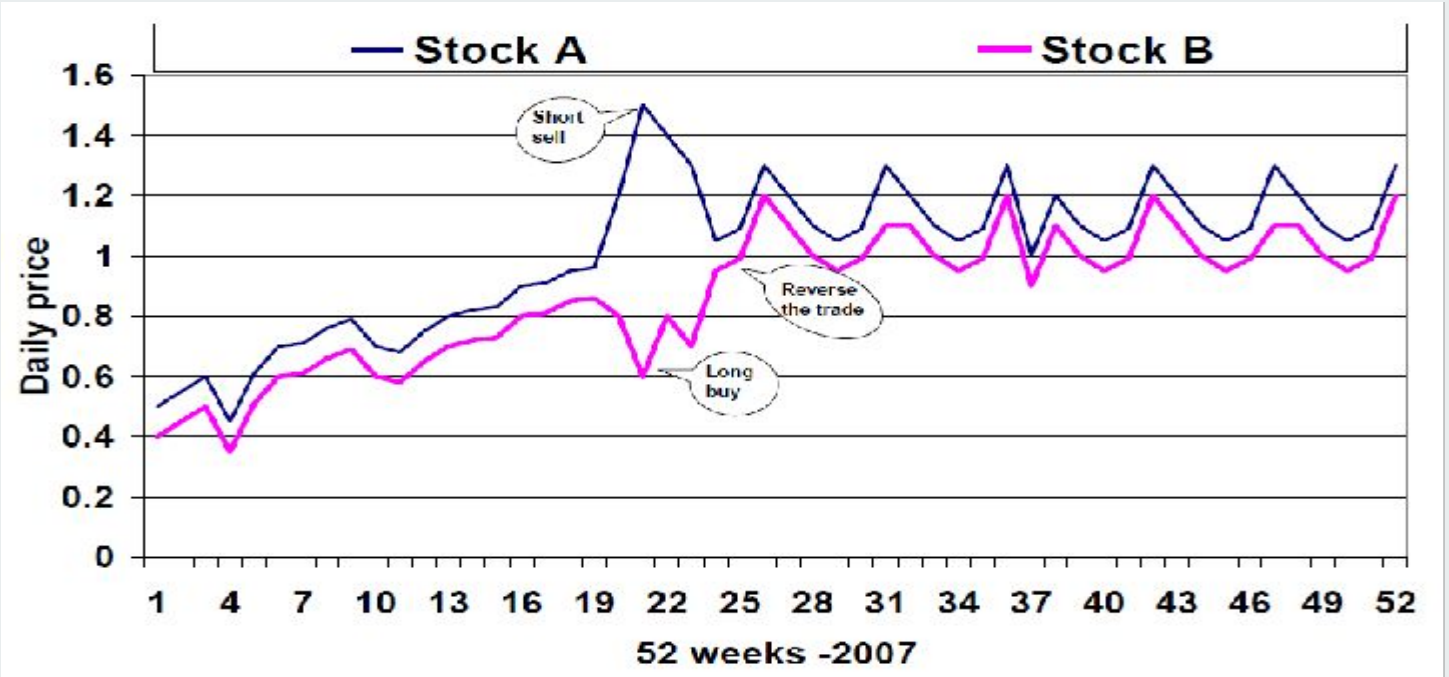
Statistical Methods in Pairs Trading

Hongyue Li

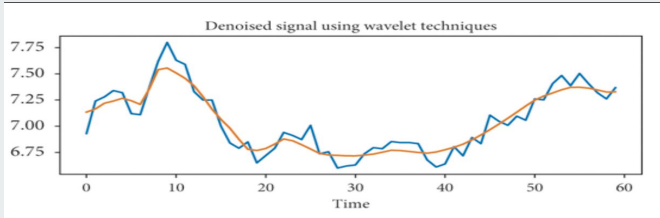


Dataset

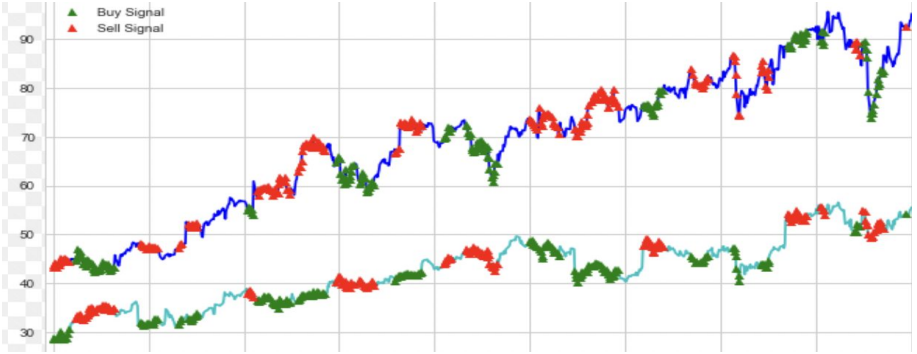
Dataset:
historical data from 2022-01-01 to 2022-12-31 of 100 popular exchange-traded funds(ETFs).
We split it into train, validation and test set by a 70:15:15 ratio.

Features: daily close, open, high, low price and volume traded for each stock.

Wavelet Denoising:
- can handle non-stationary data
- smooth out fluctuations and highlight the underlying trends
- improve the signal-to-noise ratio by reducing the noise



Pairs Trading Strategy



Method:

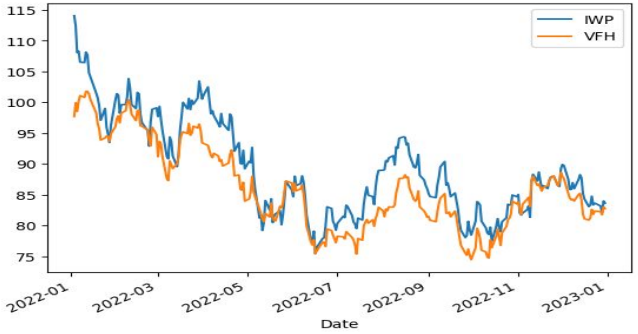
Find the pair of stocks with the highest cointegration score, and a long-short position is opened when pair prices have diverged by two standard deviations, and the position is closed when prices revert.

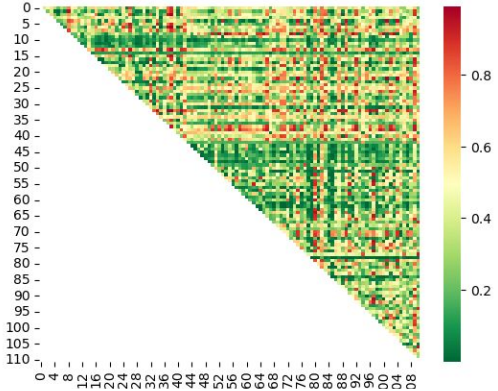
Two main tasks:

1. Identify the best pair of stocks

2. Forecast the spread of the selected pair of stocks based on historical data to make reasonable trading decisions.

Cointegration





We select the best pair of ETFs using the cointegration method. The cointegrated ETFs are expected to have the same long-term trend and exhibit mean-reverting behavior. To ensure that the identified pairs of cointegrated series is statistically significant, we calculate the p-value of the pairs and select the one with small p-value.

Based on the data, we selected the pair IWP(iShares Russell Mid-Cap Growth ETF) and VFH(Vanguard Financials Index Fund ETF) as our pair.

We check if the residual is stationary using the Augmented Dick-Fuller (ADF) test.

Pair 1	Pair 2	Value
PDP	XLF	1.027e-06
PDP	VFH	1.125e-06
PDP	PRF	1.147e-06
IWO	KRE	4.376e-06
IVV	SPY	7.271e-06
IWO	KBE	7.561e-06
PDP	RSP	2.913e-05
IWP	IYF	3.410e-05
IWP	XLF	3.879e-05
IWP	VFH	5.560e-05

Table 1: Cointegration Scores

Feature Engineering

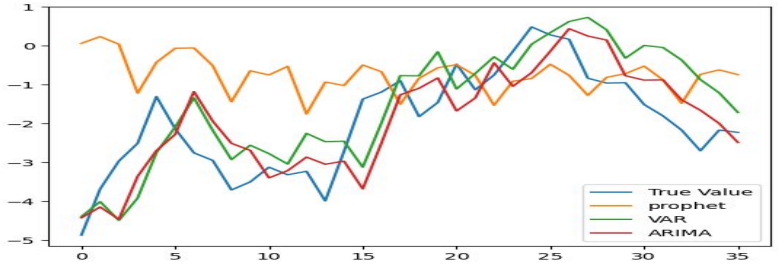
In order to create more features that can reveal the underlying dynamics of the stocks, we augmented the data with the following extracted features: relative strength index, money flow index, accumulation/distribution indicator, volume price trend, average true range, Bollinger Bands, average directional index, exponential moving average, moving average convergence/divergence, daily log return.

Relative Strength Index	Momentum (t=14)	$RSI = 100 - \frac{100}{1+RS}$ $whr\ RS = \frac{Average\ of\ t-day's\ up\ closes}{Average\ of\ t-day's\ down\ closes}$
Money Flow Index	Momentum (t=14)	$MFI = 100 - \frac{100}{MFR}$ $whr\ MFR = \frac{t-period\ Positive\ Money\ Flow}{t-period\ Negative\ Money\ Flow}$
Accumulation / Distribution Index	Volume	$ADL = ADL_{prev} + (MFM * Volume)$ $whr\ MFM = \frac{[(Close-Low)-(High-Close)]}{High-Low}$
Volume – Price Trend	Volume	$VPT = VPT_{prev} + (Volume * \frac{Close_{today} - Close_{prev}}{Close_{prev}})$
Average True Range	Volatility (n=14)	$ATR_t = \frac{ATR_{t-1}*(n-1)+TR_t}{n}$ $where\ TR_t = \max(High - Low, abs(High - Close_{prev}), abs(Low - Close_{prev}))$
Bollinger Bands	Volatility (n=20)	$BB = n-day\ Simple\ Moving\ Average$
Average Directional Movement Index	Trend (n=14)	$ADMI = Average\ of + ve\ Directional\ Movement\ and - ve\ Directional\ Movement$
Exponential Moving Average	Trend (n=14)	$EMA = [Close - EMA_{prev}] * \frac{2}{n+1} + EMA_{prev}$
Moving Average Convergence Divergence	Trend (n_fast = 14, n_slow = 30)	$MACD = EMA_{n\_fast} - EMA_{n\_slow}$
Daily Log Return	Other	$Log\ Return = ln\frac{Close_i}{Close_{i-1}}$

Classical Statistical Models

We implemented four statistical methods to predict the spread. They are Kalman Filter, ARIMA, VAR and Facebook Prophet.

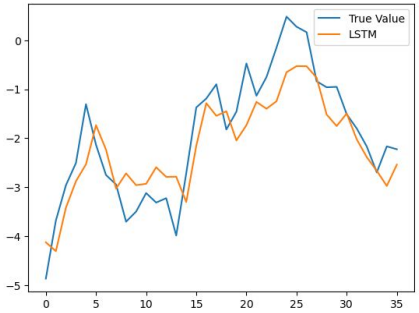
Due to the limitations of some of these models, we only use the close price as a single feature, which result in the high MSE test error for Kalman Filter and Prophet.

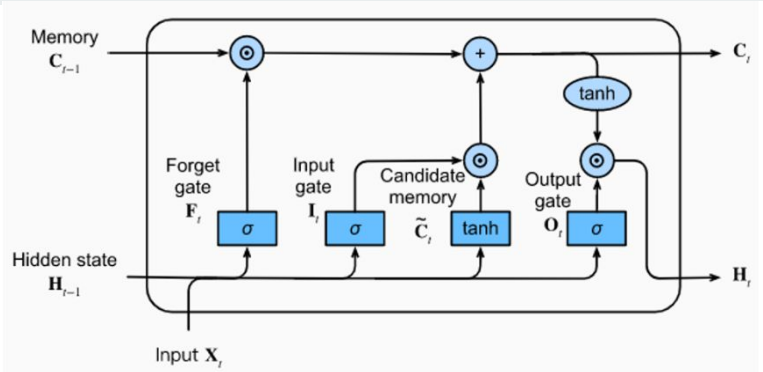


No	Method	No. Features	MSE error
1	Kalman Filter	1	4.23
2	ARIMA	1	0.83
3	VAR	All	1.16
4	Prophet	1	3.56

Table 2: Experimental Results

LSTM Model





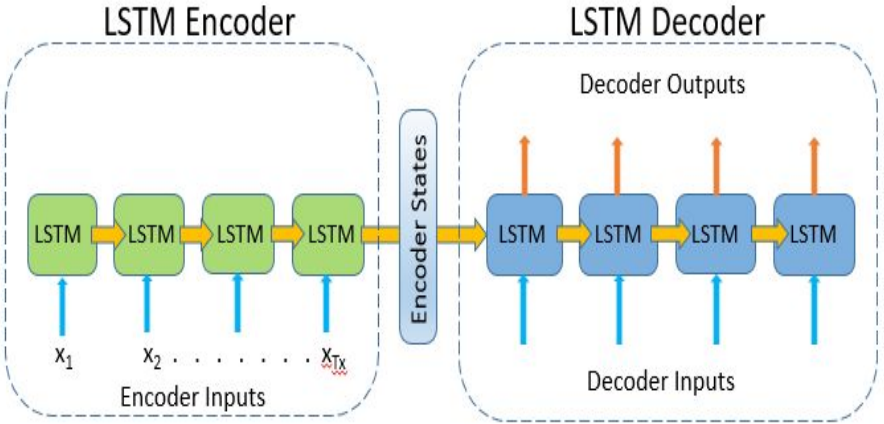
- Benefits of LSTM on time series: able to capture temporal dependencies and patterns that may not be adequately represented by traditional standalone numerical values

- Padding to same length for each time series input

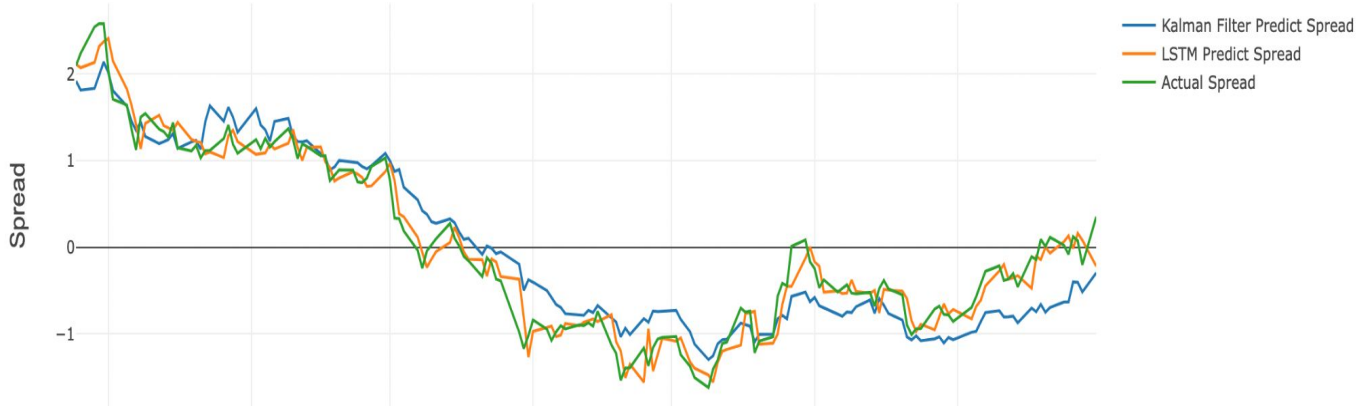
- Model Architecture: 2 LSTM Layer (hidden dim = 10) after embedding layers (hidden dim = 64). Dropout probability of 0.3 with a fully connected layers at the end.

- Possible Future directions: Attention Mechanism, Hybrid Architectures such as ConvLSTM

- Predictions: Train MSE error: 0.006 Test MSE error: 0.0156



Result



We perform a single pairs trading strategy based on the predicted spread of these models.

If the Z-score exceeds 1 or -1, we long one stock and short the other. Else if the Z-score is between -0.5 and 0.5, we clear our holdings.

The results are shown below.

Method	Expected Profit
Kalman	238.31
ARIMA	1354.46
VAR	1024.13
Prophet	-183.80
LSTM	1427.08