Mid-Frequency Statistical Arbitrage

A pairs trading strategy on securities in the Nifty500

Alistair J. Chopping

Abstract

This report accompanies the code in the corresponding GitHub repository. In this report we develop a simple example of a pairs trading strategy on two stocks in the Nifty500 at the one-minute time frequency, in an effort to demonstrate some of the core ideas of statistical arbitrage strategies. The strategy involves identifying pairs of securities that are cointegrated, namely pairs for which there exists a stationary linear combination (which we call the *spread*) of the two. Under the hypothesis that this spread will mean-revert, when it diverges from the mean one can short the overvalued security and long its undervalued counterpart. The strategy is then profitable when and if the two securities do indeed mean revert. In this report, we find that *Bajaj Finserv* and *Indian Bank* were cointegrated at the one-minute time frequency on the 26th September 2024, and we use this observation to simulate a simple pairs trading strategy.

Contents

1 Introduction

Statistical arbitrage refers to a class of short-term trading strategies that seek to exploit small inefficiencies in the prices of financial assets that are somehow related. Unlike traditional arbitrage, which involves the simultaneous buying and selling of identical assets (perhaps on different exchanges) to profit from price differences, statistical arbitrage capitalises on statistical relationships between different securities. By building a portfolio of related assets that is theoretically market-neutral, traders can potentially profit from movements in the price spreads between these assets, regardless of broader market trends.

One common strategy that falls under the umbrella of statistical arbitrage is *pairs trading*, whereby an investor identifies two assets that historically move together in some predictable way, and places complementary trades on each asset in order to capitalise on temporary deviations from their expected relationship. For instance, if two stocks are found to be cointegrated, meaning that their prices maintain a consistent, long-term equilibrium, then any divergence between them could present a trading opportunity. The strategy involves entering a long position on the undervalued stock and shorting the overvalued stock when their price spread deviates significantly from the historical mean, under the assumption that both stocks will mean revert and the trader can close out both positions for a profit.

This report focuses on implementing a pairs trading strategy using two cointegrated stocks from the NIFTY 500, an index comprised of the top 500 companies on the National Stock Exchange of India. We first narrow down our search to a small basket containing groups of stocks in similar sectors, before singling out the pair that passes a cointegration test most convincingly[1]. The pair we select is *Bajaj Finserve* and *India Bank*, which we find to be cointegrated at the one-minute frequency on the 26th of September 2024. We then perform a linear regression in order to compute the coefficient (known as the *hedge ratio*) that produces a stationary linear combination of the pair, before implementing a simplified version of a typical pairs trading strategy that captures the core concepts. Finally, we test the strategy on unseen data from the following day, and find that using the same hedge ratios, profitability of the strategy persists into 27/09/24. We conclude by mentioning a few potential improvements to the model.

2 Time Series Analysis - Cointegration & Stationarity

Two central concepts in pairs trading are those of *cointegration* and *stationarity*. We begin with a definition of a stationary time series, before defining the notions of *integration* and cointegration.

[1] Namely, we perform a hypothesis test and choose the pair that produces the lowest p-value.

Definition 2.0.1. *Let $\{X_t\}_{t\geq 0}$ be a time series, and let $F_X(x_{t_1},...,x_{t_n}),\quad n\in\mathbb{N}_{\geq 0}$ be its joint cumulative distribution function[a]. $\{X_t\}_{t\geq 0}$ is said to be **Stationary** if its joint cdf is invariant under constant shifts in time, namely*

$$F_X(x_{t_1+\tau},...,x_{t_n+\tau}) = F_X(x_{t_1},...,x_{t_n}). \tag{2.1}$$

In particular, this means that $\mathbb{E}[X_t]$ and $\mathbb{V}ar[X_t]$ are constant in time.

[a] $\{X_t\}_{t\geq 0}$ is a stochastic process like any other, namely a collection of random variables. These random variables have a joint probability distribution and therefore a joint cdf.

Definition 2.0.2. *Let $\{X_t\}_{t\geq 0}$ be a time series, and define the **lag operator** L by*

$$LX_t := X_{t-1}. \tag{2.2}$$

*X_t is said to be **integrated of order d**, or $I(d)$, if*

$$(1-L)^d X_t \tag{2.3}$$

is a stationary process.

A process is $I(d)$ if taking d repeated differences renders the process stationary. For example, an $I(0)$ process X_t is stationary by definition, since an $I(0)$ process is one for which

$$(1-L)^0 X_t = X_t, \tag{2.4}$$

is stationary.

Definition 2.0.3. *Let $\mathcal{C} = \{X_{1,t}, X_{2,t}, ..., X_{n,t}\}$ be a collection of time series which are each $I(d)$. If there exists some linear combination*

$$Z_t = a_1 X_{1,t} + ... + a_n X_{n,t}, \tag{2.5}$$

*such that Z_t is $I(m)$ with $m < d$, then the elements of the collection $\mathcal{C}$ are said to be **Cointegrated**.*

In practice, we are often concerned with the case that $d = 1$ and $m = 0$. In this report, we will be concerned only with the case that $|\mathcal{C}| = 2$.

If a pair of time series are each $I(1)$ and together they are cointegrated, namely they admit some linear combination that is $I(0)$, then as described above that linear combination will be stationary. In particular, for a cointegrated pair (X_t, Y_t) the combination

$$Z_t = Y_t - \beta X_t \tag{2.6}$$

will be stationary for some special value of β. In particular, if the mean of that stationary combination is α, we are looking for β such that

$$\alpha + \varepsilon_t = Y_t - \beta X_t \iff Y_t = \alpha + \beta X_t + \varepsilon_t, \tag{2.7}$$

where ε_t is some noise term that accounts for the random fluctuations of the stationary process. One can immediately recognise this as a linear regression problem; given observations of X_t and Y_t we can estimate β by ordinary least squares regression.

This is exactly what we do when pairs trading. We first identify a pair of stocks that are cointegrated, before running a linear regression to estimate β, which in this context is often referred to as the *hedge ratio*. In the case that we compute β for the combination $Y_t - \beta X_t$, the hedge ratio tells us how many units of asset X we should short for each unit of asset Y that we go long.

Given a random pair of stocks, how do we tell whether or not they're cointegrated? This is where we need some kind of hypothesis test. Luckily, the *statsmodels* Python library includes a method called *coint*, which performs the so-called Augemented Engle-Granger (AEG) test for cointegration. An essential part of the AEG test is in turn the Augmented Dickey-Fuller (ADF) test for stationarity, which we now describe. Though it may be considered overkill, we will also separately use the ADF test for stationarity after we have estimated the hedge ratio using linear regression, simply to be sure that the hedge ratio we have computed does indeed produce a stationary linear combination of the two series.

2.1 The (Augmented) Dickey-Fuller Test

The Augmented Dickey-Fuller test is a procedure for testing the null hypothesis that a series is non-stationary. It is an extension of the Dickey-Fuller test, whereby the starting point is the assumption that the series in question is an AR(1) [2] time series of the form

$$\begin{aligned} X_t &= \rho X_{t-1} + \varepsilon_t, \\ &= \rho^t X_0 + \sum_{k=0}^{t-1} \rho^k \varepsilon_{t-k}, \end{aligned} \tag{2.8}$$

where ε_t is a noise term with zero mean and constant standard deviation. The null and alternative hypotheses for the Dickey-Fuller test are as follows:

$$\begin{aligned} H_0 &: \rho = 1 \\ H_1 &: \rho < 1, \end{aligned}$$

which correspond to the process being non-stationary and stationary, respectively. If $\rho = 1$, the process becomes

$$X_t = X_{t-1} + \varepsilon_t, \tag{2.9}$$

and it can be shown that such a process is non-stationary; the mean will be constant but the variance will be a function of t. By contrast, if $\rho < 1$, it can be shown that the mean and variance of the process are both constant in time (with the mean in fact vanishing), and so the process is stationary[3].

If $\rho < 1$, then the series will be stationary. However, under the null hypothesis we assume that X_t is non-stationary, and therefore its joint probability distribution (or rather the parameters of

[2] An autoregressive process of order p, denoted an AR(p) process, is one of the form $X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$.

[3] We also note that $\rho > 1$ is possible, but describes "exploding" processes whose mean approaches infinity as time progresses, and thus we don't often encounter such processes in the real world.

its distribution) will vary over time. As such, we cannot immediately perform something like a t-test, since the central limit theorem will not apply[4]. Instead, let's subtract X_{t-1} from both sides, resulting in

$$\Delta X_t = \delta X_{t-1} + \varepsilon_t, \tag{2.10}$$

where we have defined

$$\Delta X_t \equiv X_t - X_{t-1}, \quad \delta \equiv \rho - 1. \tag{2.11}$$

The null and alternative hypotheses then become $\delta = 0$ and $\delta < 0$, respectively. We perform this transformation because if we now assume the null hypothesis $\delta = 0$, the LHS of the transformed process $\Delta X_t = \varepsilon_t$ is now stationary. We are better off than before, but we still can't perform a standard hypothesis test using for example the t-distribution. This is because while ΔX_t is stationary under the null hypothesis, X_t itself is not - the non-stationary δX_{t-1} term is still present in the original process (2.9) we are interested in. However, it turns out that we can still compute the t-statistic, but rather than comparing it with the t-distribution we instead compare it against a distribution tabulated by Dickey and Fuller. We compute an estimate of δ (denoted $\hat{\delta}$) with an OLS regression on (2.10), compute the t-statistic,

$$t := \frac{\hat{\delta}}{\mathrm{SE}(\hat{\delta})}, \tag{2.12}$$

with SE the standard error, and compare it against the Dickey-Fuller distribution. For $\mathrm{DF}_{\mathrm{critical}}$ the "critical value" associated with the Dickey-Fuller distribution, we reject H_0 if $t < \mathrm{DF}_{\mathrm{critical}}$, and do not reject H_0 if $t > \mathrm{DF}_{\mathrm{critical}}$.

Time series are often more complicated than AR(1), and so a natural question to ask is whether or not the Dickey-Fuller test can be extended to AR(p) time series with $p > 1$. Indeed it can, leading to the *Augmented* Dickey-Fuller test. This is essentially identical to the above; we perform a similar transformation on the series and compute the same t-statistic, before comparing to the Dickey-Fuller distribution. The only difference is that the DF test is applied to the AR(p) model

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t \longrightarrow \Delta X_t = \delta X_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta X_{t-i} + \varepsilon_t, \tag{2.13}$$

with α_i some coefficients.

3 Strategy & Results

The goal of a pairs trading strategy is to enter positions when the pair are diverging away from eachother[5], and exit positions when the pair are converging towards eachother[6]. In this section we implement such a strategy.

[4]The CLT states that for a sufficiently large sample size, the distribution of the sample mean of i.i.d random variables will be approximately normal, regardless of the original distribution of the variables. If the mean and standard deviation of the time series vary over time, then the random variables that make it up are not i.i.d, and so the CLT does not apply.

[5]Namely, when the stationary spread is moving away from its long-term mean.

[6]Namely, when the spread is moving towards its long-term mean.

3.1 Identifying Cointegrated Pairs

The code starts by loading csv files containing stock prices at the 1-minute time frequency. The folder from which they are loaded contains stocks that were hand-picked as being potentially correlated; it contains 15 stocks representing a handful of different sectors, from financial services to automobile manufacturing. By pre-selecting stocks in similar sectors, we narrow the search for cointegrated pairs down from the entire Nifty500 to a small basket of candidates. We also split the dataframes into data from 26/09/2024 and from 27/09/2024; the former of which we use for strategy development and the latter we use later on for testing purposes. Figure 1 shows a plot of the normalised[7] close prices for each stock.

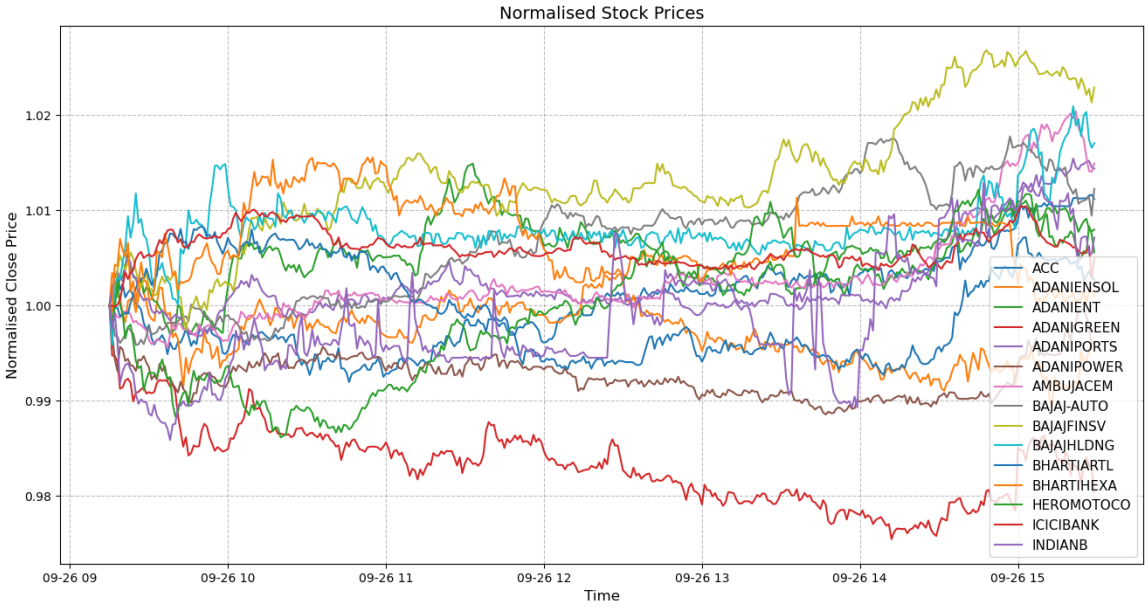

Figure 1: Normalised close prices at the one-minute frequency on 26/09/2024, for each stock in the basket of stocks identified as potentially containing correlated pairs. Various sectors are represented, from financial services to automobile manufacturing.

We then aim to identify pairs of cointegrated stocks by performing a cointegration test for every possible pair of stocks in the basket. If we find the pair to be cointegrated at the 0.01 significance level[8], we append them to a list before extracting the pair with the highest likelihood of cointegration (ie, the pair that produces the lowest p-value in the cointegration test). We find that *Bajaj Finserv* and *Indian Bank* are cointegrated with a p-value of ~ 0.001, and also have Pearson correlation $\rho = 0.8$.

[7] We normalise each series simply so that they can be displayed on the same plot.
[8] Namely, we find a p-value less than 0.01, indicating that the pair are cointegrated at a 99% level of confidence.

3.2 Finding a Stationary Spread

As described earlier, once we have identified a cointegrated pair the next step is to find a stationary linear combination of the two. We will call such a linear combination the *spread*. If we let $\mathcal{B}_t$ be the time series for Bajaj Finserv and $\mathcal{I}_t$ be that for Indian Bank, we aim to find two hedge ratios β_1 and β_2, as defined by insisting that the spreads

$$\mathcal{B}_t - \beta_2 \mathcal{I}_t \tag{3.1}$$

$$\mathcal{I}_t - \beta_1 \mathcal{B}_t \tag{3.2}$$

are stationary. β_2 tells us how much of Indian Bank to short for each unit of Bajaj Finserv that we long when the appropriate opportunity arises, and β_1 tells us the converse. We will actually trade using the series described by (3.1), but we will use the hedge ratios for both types of trade when it is appropriate. We perform two linear regressions with the function *hedge_ratios* in the code, and obtain $(\beta_1, \beta_2) \approx (0.207, 3.141)$. We then produce a dataframe describing the spread (3.1) at each time step. A final check is then made to ensure that this linear combination is indeed stationary, using the Augmented Dickey-Fuller test. Figure 2 shows a plot of the normalised close prices for the pair of stocks, together with the stationary spread showing mean-reverting behaviour. It is this mean-reverting behaviour that we will be exploiting.

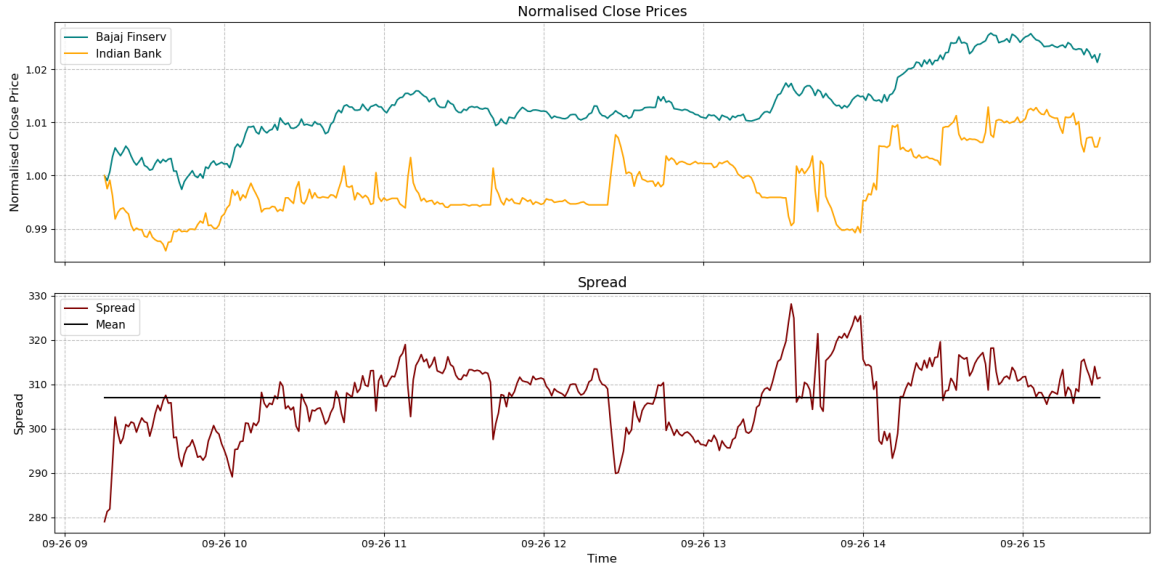

Figure 2: A plot showing the normalised close prices for Bajaj Finserv and Indian Bank, showing various periods of convergence and divergence, together with a plot of the (non-normalised) spread. Note that the spread is mean-reverting, showing increases during a period of divergence and decreases during a period of convergence.

3.3 Trade Simulation

We then define a function *strategy* to simulate trading using the hedge ratios β_1 and β_2. If the spread goes above the mean by n standard deviations, we will short β_1 units of Bajaj Finserv and long 1 unit of Indian Bank. If the spread goes below the mean by n standard deviations, we will short β_2 units of Indian Bank and long 1 unit of Bajaj Finserv. We set the strategy to have the upper and lower thresholds with a default value $n = 1$. We can then plot the cumulative returns of the portfolio over time, as shown in figure 3. On this day, the strategy has an annualised Sharpe ratio of ~ 2.6.

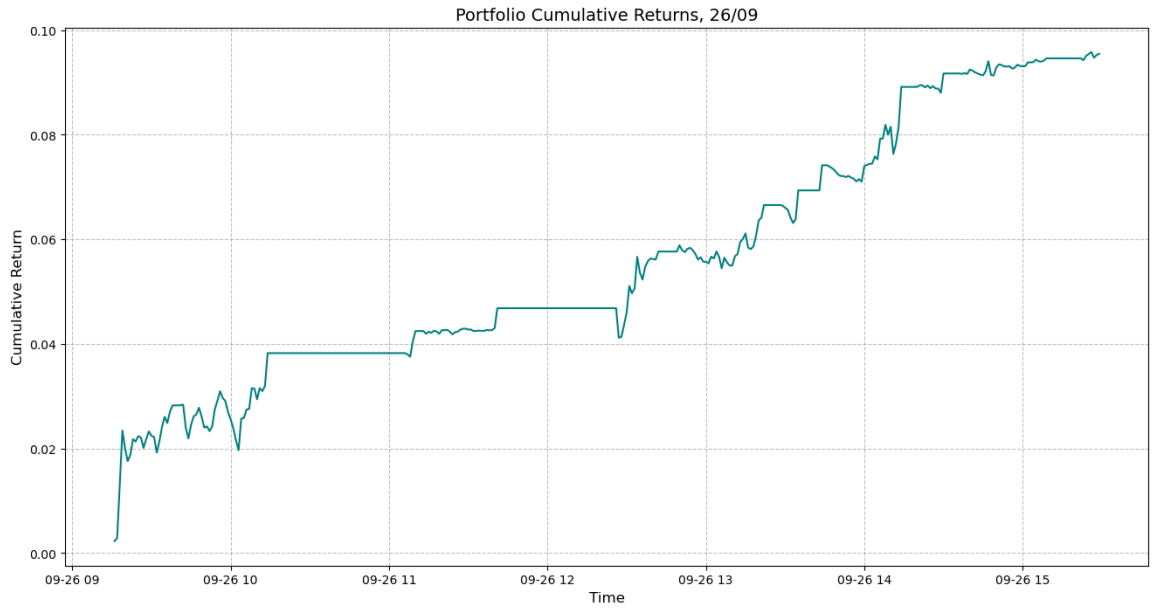

Figure 3: A plot showing the cumulative returns of the portfolio throughout the 26/09/2024 trading session. We see plateaus in the portfolio returns during periods where the spread remains close to the mean, and larger movements during larger swings in the spread. Clearly, the strategy is profitable on this day, but this is to be expected given that this is the day over which we developed the strategy. We find an annualised Sharpe ratio of ~ 2.6.

So, on 26/09/2024 the strategy makes money. This is somewhat expected, since we developed the strategy over this day. We know that the spread is stationary on this particular day using the particular β_1 we computed, we know that the two stocks are cointegrated, and we know that they remain cointegrated for the whole day. Therefore, it isn't wildly surprising that the strategy makes money. To really test whether this is a viable strategy we should test it on unseen data using these same hedge ratios; for instance on data from the following day, 27/09/2024. We don't have any information about this day; in particular we do not know whether or not the two will remain cointegrated. It seems likely that they will given that the data are only one day apart, but it isn't guaranteed.

It is straightforward to test the strategy on the following day - we simply plug the data from 27/09/2024 into the *strategy* function and plot the results. After doing so we can plot the distribution of the strategy's returns on the 27th, which is shown in figure 4.

Figure 4: A plot showing the distribution of the strategy's returns on the unseen data from 27/09/2024. We see a longer tail on the positive return side than on the negative, and more frequent small positive returns than small negative returns.

We find the strategy to also be profitable on the unseen data, despite the use of the same hedge ratios from the previous day[9]. The portfolio's cumulative returns over 27/09/2024 can be found in figure 5, and we find that the strategy's annualised Sharpe from that day is ~ 1.1.

[9]Clearly this is not the optimal approach; we will discuss improvements to the strategy in the conclusion.

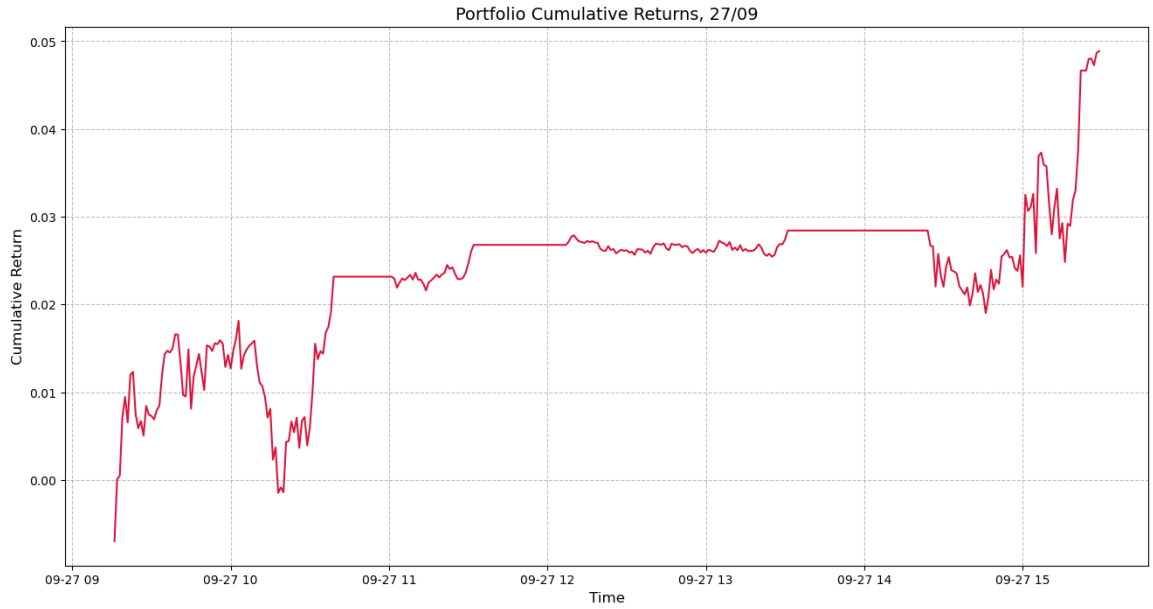

Figure 5: A plot showing the cumulative returns of the portfolio throughout the 27/09/2024 trading session. The strategy is profitable on this unseen data, though perhaps as expected, it is less profitable than on the previous day. We find an annualised Sharpe ratio of ~ 1.1.

4 Conclusion

4.1 Summary

In this report we developed a simple example of a pairs trading strategy to demonstrate some of the core ideas of statistical arbitrage strategies. We developed the strategy on data from the 26th September 2024, identifying a pair of stocks in the Nifty500 that were cointegrated at the one-minute time frequency. We employed the augmented Engle-Granger test for cointegration to identify the pair, before performing a linear regression in order to estimate the hedge ratios. We used the augmented Dickey-Fuller test to ensure that the resulting spread is stationary, before exploiting the spread's mean-reverting behaviour to enter and exit long and short positions on each asset for a profit. We showed that such a strategy is profitable on the day over which we developed the strategy, before showing that it was also profitable on unseen data from the following day, 27/09/2024. Plots can be found in figures 1, 2, 3, 4 and 5 that elucidate the strategy and give insight into its performance.

4.2 Directions for future work

- **Dynamic Hedge Ratios & Position Sizing.** In this project we computed fixed hedge ratios from data on 26/09/2024, and used these same fixed hedge ratios to test the strategy on data from 27/09/2024. Of course, this is not optimal - it would be better to dynamically update the hedge ratios throughout the day as more data comes in, in order to always ensure that we're

working with a stationary spread (under the assumption that the cointegration persists).

In the above we also employ fixed position size - we always buy and short the same amount of each stock when appropriate. As a result, the initial allocation we give the strategy is irrelevant - whether we trade with \$10 or \$10,000 we always make the same profit. Of course if one were to trade this kind of strategy in the real market, it would be prudent to adjust the position size based on external factors, such as an estimated probability of mean reversion or market conditions. This leads us into the next direction for future work.

- **Machine Learning for Predicting Market Regime.** It would be interesting to try to incorporate a machine learning model to detect market conditions, and predict for example periods of high or low volatility. We could then use this information to inform our trading decisions; for instance by placing stop losses on trades entered during periods of high volatility.
- **Trading Fees, Slippage etc.** We made no attempt in this project to account for such effects as slippage (when the price of an asset moves between order placement and fulfillment) and trading fees. Of course, were one to take such a strategy live one would need to model these to ensure that the strategy remains profitable after taking these effects into account.

  Institutional statistical arbitrage strategies typically involve not just pairs of stocks, but portfolios of hundreds of stocks that are carefully matched in order to reduce exposure to the market as a whole - such strategies are termed *beta neutral*. We made no attempt to estimate and limit the portfolio's exposure to market-wide movements in this project, and this would be an interesting extension.

Acknowledgements. I am grateful to Kaggle user Doug Barton-Smith for collecting the data and making it freely available. One can find the datasets used in this project here.