Advanced Statistical Arbitrage with Reinforcement Learning

Boming Ning[†,*] and Kiseop Lee[†]

[†]Department of Statistics, Purdue University

February 26, 2024

Abstract

Statistical arbitrage is a prevalent trading strategy which takes advantage of mean reverse property of spread of paired stocks. Studies on this strategy often rely heavily on model assumption. In this study, we introduce an innovative model-free and reinforcement learning based framework for statistical arbitrage. For the construction of mean reversion spreads, we establish an empirical reversion time metric and optimize asset coefficients by minimizing this empirical mean reversion time. In the trading phase, we employ a reinforcement learning framework to identify the optimal mean reversion strategy. Diverging from traditional mean reversion strategies that primarily focus on price deviations from a long-term mean, our methodology creatively constructs the state space to encapsulate the recent trends in price movements. Additionally, the reward function is carefully tailored to reflect the unique characteristics of mean reversion trading.

Keywords— Statistical Arbitrage, Mean Reversion Trading, Empirical Mean Reversion Time, Reinforcement Learning

1 Introduction

Statistical arbitrage, also known as mean reversion trading or pairs trading, is an important trading strategy in the financial markets. The essence of statistical arbitrage lies in creating spreads or portfolios from the market that exhibit mean-reverting characteristics, thereby unlocking opportunities for profit. For instance, if the price of a spread falls below its long-term mean, a trader might take a long position and then wait until its price correction, aiming to profit from this adjustment.

The approach to statistical arbitrage unfolds in three distinct steps: First, it entails the identification of two or more securities that have shown a historical pattern of moving together. Next, a mean-reverting spread is formulated from these correlated securities. The final step involves taking a position when the spread deviates from its long-term mean, leveraging the anticipated return to equilibrium to generate profits. Therefore, mean reversion trading is divided into three main elements: (1) the identification of securities with co-movements, (2) the construction of mean-reverting spreads, and (3) the development of a trading strategy based on these mean-reverting spreads. The first two components are referred to as the formation phase, while the third is considered the trading phase.

The initial step in statistical arbitrage strategy is the identification of similar securities. Traditional methods predominantly utilize distance metrics, as highlighted by Gatev et al. (2006), where pairs are formed by selecting the securities that minimizes the sum of squared deviations (SSD) between their normalized price series. This principle of pair selection can be extended to encompass pairs of representative stocks and ETFs within specific sectors (Gatev et al. (2006), Avellaneda and Lee (2010), Montana and Triantafyllopoulos (2011), Leung and Li (2016)), as well as physical commodities and their corresponding stocks/ETFs (Kanamura et al. (2010)) and entities within the cryptocurrency market (Leung and Nguyen (2019)). These references highlight the adaptability and efficacy of statistical arbitrage strategies across a broad spectrum of markets. In our study, we select ten representative pairs from various sectors within the US market to construct the mean reversion portfolios.

After identifying groups of similar stocks, the next step involves constructing an statistical arbitrage portfolio or spread with mean-reverting property. A foundational approach, as suggested by Gatev

*Corresponding author. Email: ningb@purdue.edu

et al. (2006), involves taking a long position in one security of the pair and a short position in the other, that is, trading on the spread $S_1 - S_2$ for two similar stocks S_1 and S_2. Although straightforward, this method often cannot create an optimal portfolio that exhibits high mean-reverting characteristics. A more sophisticated strategy frequently used is Ornstein–Uhlenbeck (OU) mean reversion trading (Leung and Li (2016)). For a pair of similar stocks S_1 and S_2, the goal is to find a coefficient B such that the spread $S_1 - B \cdot S_2$ mimics an OU process as closely as possible, with B typically determined through maximum likelihood estimation based on the OU process distribution. However, the real-world application of this strategy faces challenges due to the fact that financial markets may not always align with the assumptions of the Ornstein–Uhlenbeck process, which can compromise the effectiveness of strategies based on this model.

To overcome the limitations inherent in these assumption-dependent methods, we introduce a novel approach that utilizes the proposed empirical mean reversion time of any time series as a measure of reversion speed. This allows for the construction of a mean reversion spread without relying on any theoretical assumptions. By employing a grid search method, we can systematically explore different combinations to identify an optimal spread that possesses the least empirical mean reversion time. This technique offers a more flexible and potentially more robust framework for arbitrage portfolio construction.

The final phase of statistical arbitrage involves formulating a trading strategy based on the constructed mean-reverting spread. Traditional strategies heavily rely on model parameter estimations. For instance, Gatev et al.(2006) initiate a trade when a spread's price deviation exceeds two historical standard deviations from the mean, calculated during the pair formation phase, and exit the trade upon the next convergence of prices to the historical mean. In the context of Ornstein-Uhlenbeck (OU) mean reversion trading, parameter estimations for the long-term mean and volatility of the OU model are typically employed to define trading criteria (Leung and Li (2015), Leung and Li (2016)). These estimations depend on historical data from the formation period, with an underlying assumption that parameters remain constant in the following trading phase—an assumption that may not hold due to market fluctuations. Additionally, the selection of hyper-parameters significantly impacts trading performance, yet a robust method for the optimal hyper-parameters selection remains absent. For example, the determination of an appropriate threshold for price deviation from the mean lacks a clear consensus.

To address these challenges, we introduce a reinforcement learning (RL) algorithm designed to dynamically optimize trading decisions over time, replacing the need for predefined rules. This approach models the task within a reinforcement learning framework, aimed at enabling agents to take actions that maximize cumulative rewards in an environment. We design the state space to capture the recent movements of the spread price, thus moving away from a dependence on historical mean and standard deviation estimates. This approach enables the agent to make informed decisions about future actions by leveraging insights into current market trends, rather than depending on parameter estimations from the formation period. We get the rid of the hyper-parameters choice at the same time. Simultaneously, we remove the necessity for hyper-parameter selection by not incorporating universal hyper-parameters, such as thresholds, which can significantly impact trading performance. This approach streamlines the trading process, focusing on dynamic adaptation without the constraints of fixed parameters.

The structure of the paper is organized as follows. Section 2 provides an overview of related research in the field. Section 3 details the definition of empirical reversion time for spreads and outlines the methodology for identifying optimal asset coefficients by minimizing mean reversion time. Section 4 presents a reinforcement learning framework designed for the development of optimal trading strategies. Experimental results, based on simulated data and real-world applications in the US stock market, are discussed in Section 5. Finally, Section 6 offers conclusions and outlines future research directions.

2 Related Research

The seminal work by Gatev et al. (2006) marks a cornerstone in the study of pairs trading, a strategy predicated on the mean reversion principle. By employing what is now known as the Distance Method (DM), they analyzed CRSP stocks from 1962 to 2002, identifying trading opportunities when the price of asset pairs deviated beyond two historical standard deviations and closing positions upon price convergence. This approach yielded an excess return of 1.3 % for the top 5 pairs and 1.4% for the

top 20 pairs. Building on this, Do and Faff (2012) further examined the viability of pairs trading considering transaction costs. Their findings enhance the understanding of pairs trading's practical application, demonstrating its feasibility even when accounting for trading expenses.

Beyond the Distance Method, the stochastic spread method emerges as a significant alternative for mean reversion trading, utilizing stochastic processes to analyze the mean-reverting nature of spreads. This approach involves constructing spreads and generating trading signals based on the analysis of parameters within the chosen stochastic model. Elliott et al. (2005) were pioneers in this area, introducing a Gaussian Markov chain model to capture the mean-reverting dynamics of spreads. They leveraged model estimates against observed spread data for trading decisions, laying the groundwork for further exploration. Building on this, Do et al. (2006) expanded the concept with a generalized stochastic residual spread method, aimed at modeling relative mispricing more comprehensively. Their approach broadened the stochastic spread methodology's applicability in mean reversion trading. Further enriching this field, the extensive work by Leung and Li (2016) delves into optimal mean reversion trading strategies based on various stochastic models. Covering models such as the Ornstein-Uhlenbeck, Exponential OU, and CIR, this research illuminates the versatility and effectiveness of stochastic models in optimizing trading strategies. This progression of work significantly advances our understanding of mean reversion trading by demonstrating the potential of diverse stochastic approaches.

Cointegration tests stand as a critical alternative method for mean reversion trading strategies. Leveraging the foundational error correction model introduced by Engle and Granger (1987), Vidyamurthy (2004) outlines a cointegration framework that has become essential in pairs trading analysis. This methodology is advanced by Galenko et al. (2012), who develops active trading strategies for ETFs, utilizing cointegration to exploit trading opportunities within exchange-traded funds. Furthermore, Huck and Afawubo (2015) conducts a comparative analysis, examining the performance of the Distance Method against cointegration-based strategies within the S&P 500, clarifying the relative merits of these methodologies in the context of mean reversion trading. Demonstrating the method's extensive applicability, Leung and Nguyen (2019) crafts cointegrated cryptocurrency portfolios using both the Engle-Granger two-step approach and the Johansen cointegration test, highlighting cointegration's adaptability across different asset classes.

In recent years, the landscape of mean reversion trading has been enriched by a variety of innovative methods. Among these, the use of copulas has gained attention for its ability to model the dependence between asset pairs, as evidenced by the work of Liew and Wu (2013) and Xie et al. (2016). Additionally, an optimization approach has been explored by Zhang et al. (2020), who seek to construct sparse portfolios with mean-reverting price behaviors from multiple assets. Machine learning techniques have also emerged as a powerful tool in this domain. With contributions from Guijarro-Ordonez et al. (2021), Sarmento and Horta (2020), and Chang et al. (2021), machine learning algorithms have been demonstrated to be able to uncover complex statistical patterns and relationships among assets. These developments signal a period of significant innovation in statistical arbitrage, providing traders and researchers with an expanded toolkit for strategy development and implementation.

3 Empirical Mean Reversion Time: Spread Construction

Once we identify groups of similar stocks, we need to form an arbitrage portfolio with mean reversion property. In the traditional OU pairs trading, for two similar stocks S_1 and S_2, we choose B such that the spread $S_1 - BS_2$ follows an OU process as closely as possible. A reasonable way to find B is to use the maximum likelihood estimator, using the distribution of the OU process. However, since we do not have a model assumption, we cannot use MLE anymore.

We extend paired trading to a multi-asset portfolio. In other words, given n similar stocks $S_i, i = 1, 2, ..., n$, we form a spread $X = \sum_{i=1}^{n} a_i S_i$. Our goal is to find a portfolio $(a_1, a_2, ..., a_n)$ such that the spread Y has a mean reverting property as much as possible.

Let us consider a popular OU process

$$dX_t = \mu(\theta - X_t)dt + \sigma dW_t,$$

where W_t is a standard Brownian motion. An empirical result shows that μ has the biggest impact on the profit among three parameters μ, θ and σ. In general, a larger μ gives higher return. Intuitively, it should be the case, since a larger μ implies a faster mean reversion. Therefore, a trader can make a quick profit by taking advantage of the deviation from the mean.

Therefore, we want to make the spread $X = \sum_{i=1}^{n} a_i S_i$ to have a faster mean reversion property. Consider a trading strategy to buy at $X_t = \theta - a$ and sell later at $X_t = \theta$ for $a > 0$ and a given long-term mean θ. Define the stopping time

$$\tau_t = \inf\{s > t : X_s = \theta \mid X_t = \theta - a\}. \tag{1}$$

A faster mean reversion corresponds to a smaller τ_t. Based on this logic, it is natural to define our arbitrage portfolio selection problem as follows.

At time t, consider the training time interval $[t-h, t]$. Find the optimal portfolio $(a_1, a_2, ..., a_n)$ which minimizes the sample mean of τ's in this interval, given $\bar{Y} = \theta$ and $S^2(Y) < M$ for a constant M.

The reason we impose the upper bound M on the sample variance is because of the constraint of the initial wealth and to prevent a large leverage, which makes the portfolio unstable.

3.1 Empirical Mean Reversion Time

In this part, we introduce the conception of empirical mean reversion time based on the idea above.

Inspired by Fink and Gandhi (2007), we firstly define the important extremes of time series. Let s be the sample standard deviation of a time series X_t, where $t \in [0, T]$. In real market, we can only get the discrete data points. So we assume a time series $(X_1, \cdots, X_n)$. Let C be a positive constant. A point X_m is an important minimum of the time series if there are indices i and j, where $i \leq m \leq j$, such that

- X_m is the minimum among $X_i, \cdots, X_j$;
- $X_i - X_m \geq C \cdot s$ and $X_j - X_m \geq C \cdot s$.

Intuitively, X_m is the minimal value of some segment $X_i, \cdots, X_j$, and the endpoint values of this segment are much larger than X_m. Similarly, X_m is an important maximum if there are indices i and j, where $i \leq m \leq j$, such that

- X_m is the maximum among $X_i, \cdots, X_j$;
- $X_m - X_i \geq C \cdot s$ and $X_m - X_j \geq C \cdot s$.

Drawing on the conceptual framework introduced by Equation (1), our objective is to propose an empirical mean reversion time that quantifies the duration required for the spread to revert to its long-term mean, starting from the maximum deviation observed. It enables us to infer the optimal coefficients for securities by minimizing the spread's empirical reversion time.

We now proceed to construct a sequence of time moments $\{\tau_i\}_{i=0}^{N}$, derived recursively from the significant local extremes within the actual asset price process. More precisely, we define the initial time moment as

$$\tau_1 = \inf\{u \in [0, T] : X_u \text{ is a local extreme}\}.$$

Subsequently, τ_2 is identified as the first instance when the series crosses the sample mean $\hat{\theta}$, defined by

$$\tau_2 = \inf\{u \in [\tau_1, T] : X_u = \hat{\theta}\}.$$

Recursively, τ_3 is the first local extreme following τ_2, and τ_4 is the first crossing of the long-term mean after τ_3, and so on. Thus, all odd-numbered time moments $\{\tau_n\}_{n=1,3,5,\cdots}$ correspond to local extremes and are defined as

$$\tau_n = \inf\{u \in [\tau_{n-1}, T] : X_u \text{ is a local maximum}\}.$$

Conversely, all even-numbered time moments $\{\tau_n\}_{n=2,4,6,\cdots}$ are associated with the crossings of the long-term mean, specified by

$$\tau_n = \inf\{u \in [\tau_{n-1}, T] : X_u = \hat{\theta}\}.$$

The complete sequence $\{\tau_n\}_{n=1}^{N}$ is constructed in an inductive manner.

Once we get the time stamps of iterated time stamps $\{\tau_n\}$, the empirical reversion time r is defined as the average of the time interval from local extremes to crossing times. That is,

$$r = \frac{2}{N} \sum_{\substack{i=2 \\ i \text{ even}}}^{N} (\tau_n - \tau_{n-1})$$

Next, we briefly introduce a *grid search algorithm* that can help us find the optimal coefficients based on the empirical mean reversion time. Assume the price processes of n similar assets are denoted by $S_1, S_2, \ldots, S_n$. Our aim is to find the optimal coefficients $(a_1, a_2, \ldots, a_n)$ such that the portfolio $X = \sum_{i=1}^{n} a_i S_i$ exhibits the minimal empirical mean reversion time. Without loss of generality, we set the first coefficient to $a_1 = 1$. We then evaluate the empirical mean-reversion time of Y for each coefficient a_i, where $a_i \in [-3.00, -2.99, -0.98, \ldots, 2.99, 3.00]$ for $2 \leq i \leq N$. The optimal coefficients are determined by selecting the set that minimizes the empirical mean reversion time of Y.

4 Reinforcement Learning: Advanced Trading Strategies

The final stage of statistical arbitrage involves developing a trading strategy based on a mean-reverting spread. Traditional approaches assume parameters' stability from formation phase to trading phase, which market changes can challenge. Moreover, the choice of hyper-parameters, such as the deviation threshold, critically affects performance, yet a standard method for their optimal selection is lacking.

Our motivation is to leverage reinforcement learning algorithms to help us decide the optimal trading actions dynamically over time, other than design some preset rules manually. Reinforcement learning framework is a machine learning method concerned with how intelligent agents ought to take optimal actions in an environment in order to maximize the cumulative reward.

4.1 Preliminaries of Reinforcement Learning

In RL, the sequential decision-making problem is modeled as Markov decision process (MDP), which is an augmented structure of Markov process. In addition to a Markov process, one has the possibility of choosing an action from an available action space and get some reward that tells us how good our choices were at each step.

The "environment" is defined as the part of the system outside of the RL agent's control. At each time step t, we observe the current state of the environment $S_t \in \mathcal{S}$ and then chooses an action $A_t \in \mathcal{A}$. The choice of action influences both the transition to the next state, as well as the reward received, R_t. Thus, we will get a sequence in MDP as:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \cdots$$

Every MDP is uniquely determined by a multivariate conditional probability distribution $p(s', r|s, a)$, which is the joint probability of transitioning to state s' and receiving reward r, conditional on the previous state being s and taking action a.

A policy π is a mapping from states to probability distributions over the action space. If the RL agent is following policy π, then in state s it will choose action a with probability $\pi(a|s)$. To find the optimal policy, one must specify a goal function. A wide-used discounted goal function is defined as

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \\ &= R_{t+1} + \gamma G_{t+1}, \end{aligned} \tag{2}$$

where R_t is the instant reward at time t and $\gamma \in (0, 1)$ is a discount factor expressing that rewards further in the future are worth less than rewards which are closer in time. Our goal is to search for the optimal policy that maximizes the expectation of the goal function, namely

$$\max_{\pi} \mathbb{E}[G_t]$$

Next we introduce related concepts of Q-learning. The action-value function for policy π is the expectation of goal function, assuming we start in state s, take action a and then follow the policy π from then on

$$q_\pi(s, a) := E[G_t|S_t = s, A_t = a].$$

The optimal action-value function is then defined as

$$q_*(s, a) = \max_{\pi} q_\pi(s, a).$$

If we knew the optimal action-value function, we would know the optimal policy itself easily, that is, choose $a \in \mathcal{A}$ to maximize $q_*(s, a)$. Hence we can reduce the problem to finding q_*, which is solved

iteratively based on the Bellman equations. It is straightforward to establish the Bellman equation of action-value function:

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma \max_{a'} q_*(s',a')]. \tag{3}$$

The core of the Q-learning is to leverage Bellman equation as a simple value iteration update, using the weighted average of the old value and the new information.

Before learning begins, the approximate action-value function Q is initialized to a possibly arbitrary value. Then, the corresponding sample update for q-function of S_t, A_t, given a sample next state and instant reward, S_{t+1} and R_{t+1} (from the model), is the Q-learning update:

$$Q^{new}(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \cdot \left(R_{t+1} + \gamma \cdot \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right), \tag{4}$$

where α is the learning rate, γ is the discount factor, R_{t+1} is the reward received after taking action A_t in state S_t, and S_{t+1} is the next state. Note that $Q^{new}(S_t, A_t)$ is the sum of three factors:

1. $(1-\alpha)Q(S_t, A_t)$: the current value weighted by the learning rate.
2. $\alpha \cdot R_{t+1}$: the weighted instant reward to obtain if action A_t is taken when in state S_t.
3. $\alpha \cdot \gamma \cdot \max_a Q(S_{t+1}, a)$: the maximum cumulative reward that can be obtained from next state S_{t+1} (weighted by learning rate and discount factor)

Generally, $R_{t+1} + \gamma \cdot \max_a Q(S_{t+1}, a)$ is referred as target Y_t. Thus, the iteration (4) updates the current value $Q(S_t, A_t)$ towards a target value Y_t.

An epsilon-greedy strategy is employed for action selection. In the training phase, actions are chosen at random with a probability of ϵ, whereas the action with the highest Q-value is selected with a probability of $1-\epsilon$. This approach facilitates a balance between exploration of new actions and exploitation of known values. In the testing phase, when the trained agent is assessed using new incoming data, ϵ is adjusted to 0. This modification ensures that action selection is solely based on the highest Q-value, thereby focusing entirely on exploitation based on the acquired knowledge. Thus, the model incorporates both exploration and exploitation during training, while adopting a strategy of pure exploitation during testing.

4.2 RL Model for Mean Reversion Trading

Now we can introduce our reinforcement learning model for an optimal mean reversion trading strategy.

The state space is constructed based on the trajectory of price movements over a recent sequence of time points. At any particular moment t, the state, denoted S_t, is encapsulated by the vector

$$S_t = [d_{t-l+1}, d_{t-l+2}, \ldots, d_t],$$

where each d_i characterizes the direction and magnitude of price changes at time i. A positive value of d_i signifies a price increase relative to time $i-1$, and conversely, a negative value indicates a decline. The magnitude of d_i quantifies the extent of this change. Formally, for each i within the interval $t-l+1 \le i \le t$, let $\pi_i = \left(\frac{P_i - P_{i-1}}{P_{i-1}}\right) \times 100$ represent the percentage price change from $i-1$ to i. The definition of states is given by

$$d_i = \begin{cases} +2 & \text{if } \pi_i > k, \\ +1 & \text{if } 0 < \pi_i < k, \\ -1 & \text{if } -k < \pi_i < 0, \\ -2 & \text{if } \pi_i < -k. \end{cases}$$

Here, '+2' indicates a significant increase, '+1' a moderate increase, '-2' a significant decrease, and '-1' a moderate decrease. The choice of threshold k, such as 3%, is adjustable to accommodate different sensitivity levels. This approach results in a state space comprising 4^l unique states, effectively capturing a wide spectrum of recent price movement scenarios.

The action space is composed of three possible actions: selling one share is represented by -1, taking no action is denoted by 0, and buying one share is indicated by $+1$. The set of available actions at any given time is contingent upon the agent's current position. Specifically, when the agent does

not hold any position, the permissible actions include buying (+1) or holding (0). Conversely, if the agent is currently in a long position, the options are limited to selling (−1) or holding (0). Note that our model does not account for initially entering a short position.

The immediate reward, R_{t+1}, earned by the agent for taking action A_t under prevailing environmental conditions, is mathematically defined as:

$$R_{t+1} = A_t \cdot (\theta - X_t) - c \cdot |A_t|, \tag{5}$$

where X_t denotes the current price of the spread, and θ represents the true global mean of X_t. The formulation is designed such that a buy action ($A_t = +1$) is rewarded positively when X_t is below its long-term mean, θ, encouraging purchases at lower prices. Conversely, a sell action ($A_t = -1$) incurs a negative reward under the same conditions. If X_t exceeds the long-term mean, resulting in a negative value for $\theta - X_t$, the rewards for buy and sell actions are adjusted accordingly to discourage buying at high prices and encourage selling. The term c represents the transaction cost per trade.

The cumulative return from time t to the terminal time T is expressed as:

$$G_t = \sum_{s=t+1}^{T} e^{-r\cdot(s-t)} \cdot R_s + I_T \cdot X_T, \tag{6}$$

where r denotes the interest rate, reflecting the time value of money, and I_T signifies the position held at the terminal time. This formulation accounts for the exponential decay of rewards over time due to the discounting effect of the interest rate, emphasizing the importance of immediate gains and the impact of holding a position until the end of the considered period.

To accurately fit the optimal Q-table, ample training data is essential. However, the real market offers only a limited observation path of spreads, presenting a significant challenge for effective training. Additionally, our reward function, as defined in Equation (5), incorporates the true long-term mean of the spread—a value that remains elusive in actual market scenarios.

To overcome these obstacles, our strategy involves initially simulating a multitude of mean reversion spreads with different parameters to train the reinforcement learning (RL) agent. This step allows for extensive exposure to various market conditions, enhancing the agent's learning and decision-making capabilities. Subsequently, the model, now adept from the simulation training, is applied to execute trades in the real market. In the language of RL, this method entails utilizing a simulated environment for the agent's training phase.

5 Experiments

In this chapter, the performance of the proposed method is thoroughly evaluated. Initially, tests are conducted on simulated data to assess the efficacy of the proposed empirical mean reversion time, and the reinforcement learning (RL) model. Subsequently, mean reversion trading experiments are carried out on the S&P 500 using the proposed model-free framework, with outcomes compared against those achieved with other classical statistical arbitrage methods.

5.1 Empirical Mean Reversion Time

In this section, we investigate the empirical mean reversion time by conducting simulations of the Ornstein-Uhlenbeck (OU) process. We fix the parameters $\theta = 0$ and $\sigma = 1$, and vary μ from 2 to 20 in increments of 2. For each parameter combination, we simulate 100 paths of the OU process, each with a terminal time of $T = 1.0$ and $n = 1000$ data points. Subsequently, we calculate the average empirical mean reversion time using a threshold of $C = 2$ for these paths. The primary objective is to compute the average empirical mean reversion time for these paths using a threshold value of $C = 2$, thereby examining the impact of the mean reversion parameter μ on the empirical mean reversion time.

Figure 1 presents the identified local extremes on a simulated Ornstein-Uhlenbeck (OU) path characterized by parameters $\mu = 10$, $\theta = 0$, and $\sigma = 1$. The analysis reveals that the defined criteria for extreme points are highly effective in pinpointing nearly every instance where the time series reaches a local maximum or minimum. This capability underscores the precision of our approach in capturing significant turning points within the simulated path, providing a robust method for analyzing mean

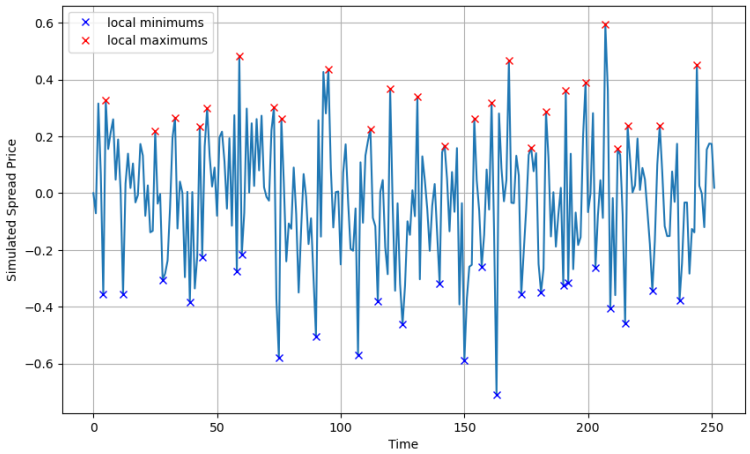


Figure 1: Local extremes calculated on a simulated OU spread with $\theta = 10$, $\theta = 0$ and $\sigma = 1$.

Parameter μ	Average EMRT	Parameter μ	Average EMRT
2.0	98.79	12.0	49.22
4.0	83.45	14.0	45.10
6.0	78.09	16.0	38.04
8.0	59.22	18.0	35.63
10.0	58.51	20.0	31.15

Table 1: Variation of average empirical mean reversion time (EMRT) with parameter μ in the Ornstein-Uhlenbeck Process.

reversion characteristics. The accurate identification of these extremes is critical for the development of mean reversion time.

The results of the impact of the mean reversion parameter μ on the empirical mean reversion time are succinctly presented in Table 1, which supports our initial hypothesis by demonstrating a clear inverse relationship between the mean reversion speed, μ, and the empirical mean reversion time. As μ increases, the mean reversion time decreases, indicating a faster adjustment of the process back to its mean level. This finding confirms our hypothesis that the empirical mean reversion time reflects the mean-reverting speed of financial time series.

5.2 RL Trading on Simulated Data

In this part, we introduce a preliminary simulated experiment into the effectiveness of the proposed reinforcement learning strategy tailored for mean reversion trading. With the parameters fixed at $\mu = 1$, $\theta = 1$, and $\sigma = 0.1$, we simulate 10,000 paths of the Ornstein-Uhlenbeck process. Each path is designed to reach a terminal time of $T = 252$ and includes $n = 252$ data points, effectively simulating one year of data for use as training samples in our study. This process was chosen due to its relevance in modeling mean-reverting financial instruments, thereby providing a realistic and challenging environment for training our reinforcement learning model.

Our reinforcement learning (RL) model configuration employs a lookback window size of $l = 4$, generating 16 distinct states. Hyper-parameters are set with a learning rate of 0.1, a discount factor of 0.99, an epsilon of 0.1 for the epsilon-greedy strategy and 10 training episodes.

Following the training phase, we apply the trained model to a new, distinct OU sample path to evaluate its decision-making power. This simulated trading are visually presented in Figure 2, where buy and sell actions executed by the RL agent are denoted by green and red points, respectively. The outcome demonstrates that the trained agent is capable of executing a series of strategic buy and sell decisions.

Subsequently, we evaluate the trained reinforcement learning model on 100 new samples, calculating the average accumulated profits across these samples. An initial investment of 100 dollars is allocated

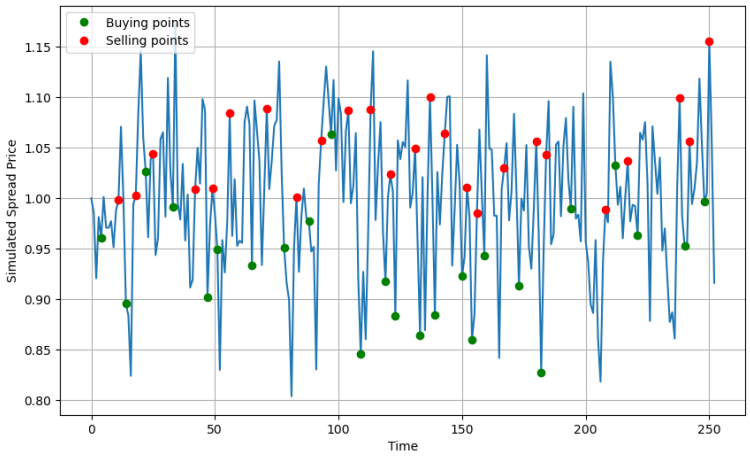


Figure 2: Simulated trading on OU process based on reinforment learning framework.

to each new OU sample. At each purchase point recommended by the RL agent, the entire available cash is used to take a long position on the spread. These positions are then closed at the selling points suggested by the agent. The simulation results in an average profit exceeding 600% across these 100 paths, underscoring the model's adeptness at identifying and leveraging trading opportunities following its training period.

5.3 Real World Experiments

In this section, we conduct real-world experiments on S&P500 to evaluate the performance of our proposed strategy, comparing it against established benchmarks such as the classic distance method (DM) [9] and the Ornstein-Uhlenbeck (OU) mean reversion trading strategy [1, 14].

5.3.1 Benchmarks

We begin by introducing the details of the implementation of these benchmark strategies.

For the distance method [9], the initial step involves calculating the sum of squared deviations among all potential pairs' normalized price series. This is followed by the identification and selection of pairs of securities that yield the minimum sum of squared deviations. Subsequently, a mean-reverting spread is constructed, denoted as $X = S_1 - S_2$, where S_1 and S_2 represent two analogous stocks. In this context, a long position is assumed in one security of the pair, while a short position is taken in the other. Additionally, estimates of the long-term mean and standard deviations are determined during the formation period.

Transitioning to the trading phase of distance method, the strategy prescribes initiating a long position when the spread's price deviation falls below multiples of estimated standard deviations from the long-term mean. The trade is then exited upon the subsequent reversion of prices. To be more clear, we clarify the trading criterion that we use in our experiment:

- buy to open if $X_t - \bar{x} < -k \cdot s$
- close long position if $X_t - \bar{x} > k \cdot s$

where $\bar{x}$ and s are the sample mean and standard deviance estimated from the formation period. The threshold parameter k is set to 1 in our experiment, Note that we solely focuses on the initial engagement in a long position within the portfolio.

For the OU mean reversion trading strategy [1, 14], we also select the pairs of securities that yield the minimum sum of squared deviations. The next step consists of constructing mean-reverting spreads for the pairs, denoted as $X = S_1 - B \cdot S_2$, where S_1 and S_2 represent two analogous stocks and B is determined by maximizing the likelihood score of fitting the spread to an OU process. Furthermore, the parameters of the spreads are estimated as an OU process, including the mean reversion speed $\hat{\mu}$, the long-term mean $\hat{\theta}$, and the volatility $\hat{\sigma}$.

Pairs Index	Pairs Trading Coefficient B		
	DM	OU	EMRT
MSFT-GOOGL	1.0	0.99	0.89
CVS-JNJ	1.0	0.43	-0.24
CL-KMB	1.0	0.39	0.46
V-MA	1.0	0.53	0.33
GE-BA	1.0	0.20	0.34
OXY-XOM	1.0	0.77	0.22
WELL-VTR	1.0	0.99	0.98
PPG-SHW	1.0	0.33	0.12
VZ-TMUS	1.0	0.10	0.01
CSX-NSC	1.0	0.12	0.14

Table 2: Comparison of pairs coefficients derived from various methods.

In the OU trading phase, the equilibrium variance is calculated as $\hat{\sigma}_{eq} = \frac{\hat{\sigma}}{\sqrt{2\hat{\mu}}}$, according to Avellaneda and Lee (2010). The basic trading signals are based on the estimations of the OU parameters:

- buy to open if $X_t - \hat{\theta} < -k \cdot \hat{\sigma}_{eq}$
- close long position if $X_t - \hat{\theta} > k \cdot \hat{\sigma}_{eq}$

where k represents the cutoff value and we set it to 0.5 in our experiment. The trading remains exclusively on initially entering a long position in the portfolio.

5.3.2 Data

Our experimental framework is designed to encompass a one-year formation period, subsequently followed by a trading period spanning the subsequent year. We utilize the daily adjusted closing prices of representative stocks from different sectors within the U.S. market to construct mean reversion spreads. Our selection includes pairs such as MSFT-GOOGL from Technology, CVS-JNJ from Healthcare, CL-KMB from Consumer Goods, V-MA from Financials, GE-BA from Industrials, OXY-XOM from Energy, WELL-VTR from Real Estate, PPG-SHW from Materials, VZ-TMUS from Telecommunication, and CSX-NSC from Transportation. Data on the daily closing prices for these stocks was collected over the period from January 1, 2022, to December 31, 2023. The data was sourced from the Yahoo! Finance API[1].

Following the completion of the annual trading, we will collate and analyze the data to calculate the trading performance across various sectors. This process is designed to rigorously evaluate the efficacy of our strategy across different market segments, thereby demonstrating its consistency and adaptability in the face of financial market uncertainties.

5.3.3 Experimental Results

In the forthcoming part, we delve into a detailed analysis of a one-year study, with 2022 designated as the formation period and 2023 as the trading period. During the formation phase, we construct mean reversion portfolios denoted as $X = S_1 - B \cdot S_2$, where S_1 and S_2 symbolize the first and second stocks, respectively, as listed above. Here, B represents a positive coefficient tailored to each trading strategy. Specifically, for the Distance Method, this coefficient is uniformly set to 1 across all pairs. In contrast, for OU pairs trading, B is determined by optimizing the likelihood score of the pair's fit to an Ornstein-Uhlenbeck (OU) process. Within our proposed methodology, B is calibrated by aiming to minimize the empirical mean reversion time of the spread. Table 2 compiles the pairs trading coefficient B for each selected pair's mean reversion portfolio, comparing the benchmarks with our novel method.

Figure 3 presents the evolution of total wealth throughout the year 2023, offering an initial comparison of trading performance between the proposed method and established baselines. This visualization provides a preliminary insight into the efficacy of our strategy relative to conventional benchmarks.

[1] https://pypi.org/project/yfinance/

Index	MSFT-GOOGL	CVS-JNJ	CL-KMB	V-MA	GE-BA
			DM Method		
DailyRet (%)	0.0446	0.0440	0.0659	-0.0244	0.2771
DailyStd (%)	0.4670	2.0692	1.2314	1.1796	3.9356
DailySR	0.0955	0.0213	0.0535	-0.0207	0.0704
MaxDD (%)	-2.1344	-24.6778	-13.6791	-10.7238	-18.1823
CumulPnL (%)	11.4443	5.7581	15.6385	-7.4888	64.2387
			OU Method		
DailyRet (%)	0.0327	-0.0073	0.0198	0.0342	0.0392
DailyStd (%)	0.4285	1.4950	0.7589	0.3475	0.3890
DailySR	0.0764	-0.0049	0.0261	0.0985	0.1007
MaxDD (%)	-2.1427	-25.6665	-5.7253	-1.0185	0.000
CumulPnL (%)	8.2443	-4.5179	4.298	8.7348	10.046
			RL Method		
DailyRet (%)	0.1344	0.0585	0.0826	0.0330	0.1679
DailyStd (%)	1.0754	0.7506	0.6000	0.3144	1.2803
DailySR	0.1250	0.0780	0.1377	0.1049	0.1312
MaxDD (%)	0.0000	0.0000	-1.9476	-0.6211	0.0000
CumulPnL (%)	37.7555	14.8895	22.2879	8.4248	48.8196

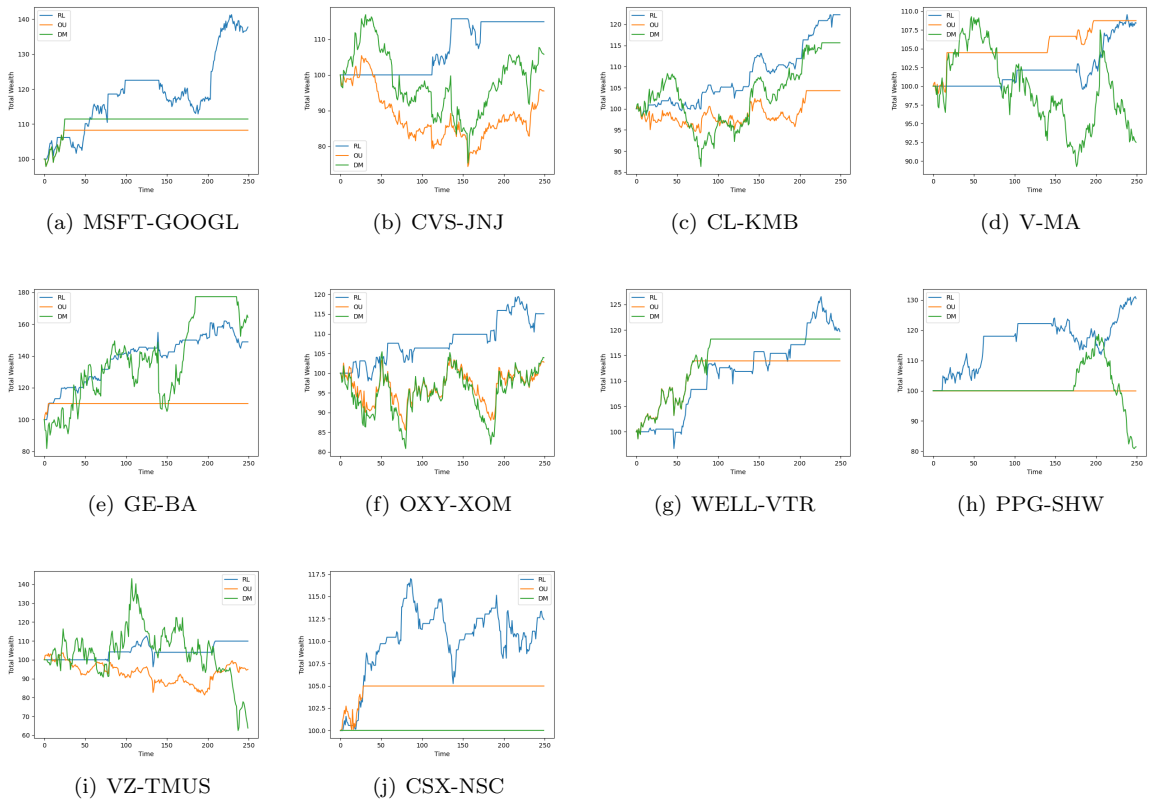
Table 3: Performance summary for trading mean reversion portfolios by baselines and the proposed RL method.

The trading performance of our proposed method in comparison to established benchmarks is detailed in Tables 3 and 4. We present a comprehensive set of performance metrics including daily returns (DailyRet), daily standard deviation (DailyStd), daily Sharpe Ratio (DailySR), maximum drawdown (MaxDD), and the annual cumulative profit and loss (CumulPnL).

Our experimental results demonstrate that the proposed reinforcement learning approach significantly outperforms traditional benchmarks in terms of daily Sharpe Ratio and cumulative returns, thereby evidencing its effectiveness and robustness in executing mean reversion trading strategies across diverse market sectors. This distinct out-performance underscores the potential benefits of incorporating reinforcement learning techniques into mean reversion trading frameworks. A pivotal factor in achieving such success is the careful design of the reinforcement learning framework, tailored to align with the specific nuances and challenges of the financial applications.

Index	OXY-XOM	WELL-VTR	PPG-SHW	VZ-TMUS	CSX-NSC
	DM Method				
DailyRet (%)	0.0373	0.0694	-0.0772	-0.112	0.0000
DailyStd (%)	2.0950	0.6555	1.0113	3.7311	0.0000
DailySR	0.0178	0.1058	-0.0764	-0.0300	0.0000
MaxDD (%)	-19.1535	-1.4004	-19.1196	-37.4779	0.0000
CumulPnL (%)	3.9194	18.2114	-18.5547	-36.1756	0.0000
	OU Method				
DailyRet (%)	0.0238	0.0539	0.0000	-0.0123	0.0199
DailyStd (%)	1.6012	0.5480	0.0000	1.3241	0.2879
DailySR	0.0149	0.0983	0.0000	-0.0093	0.0693
MaxDD (%)	-14.5001	-1.3798	0.0000	-18.652	0.0000
CumulPnL (%)	2.7812	13.9245	0.0000	-5.0869	4.9825
	RL Method				
DailyRet (%)	0.0609	0.0745	0.1124	0.0412	0.0496
DailyStd (%)	0.9446	0.6794	1.0600	0.8037	0.7101
DailySR	0.0861	0.1097	0.1061	0.0513	0.0698
MaxDD (%)	-2.0008	-3.2895	0.0000	-3.7306	0.0000
CumulPnL (%)	15.0791	19.6910	30.4559	9.9163	12.4263

Table 4: Performance summary for trading mean reversion portfolios by baselines and the proposed RL method.



(a) MSFT-GOOGL (b) CVS-JNJ (c) CL-KMB (d) V-MA

(e) GE-BA (f) OXY-XOM (g) WELL-VTR (h) PPG-SHW

(i) VZ-TMUS (j) CSX-NSC

Figure 3: 2023 total wealth growth for trading mean-reverting portfolios based on benchmarks and the proposed RL method. The initial investment is $100.

6 Conclusions and Future plan

This study has presented a novel approach to statistical arbitrage by integrating a model-free framework with reinforcement learning techniques. By establishing an empirical mean reversion time metric and optimizing asset coefficients to minimize this duration, our work has substantially refined the process of constructing mean reversion spreads. Furthermore, we have formulated a reinforcement learning framework for the trading phase, carefully designing the state space to encapsulate the recent trends in price movements and the reward functions to align with the distinct attributes of mean reversion trading. The empirical analysis conducted over the several sectors within the US market has underscored the proposed method's effectiveness and consistency.

For future work, we aim to explore more sophisticated reinforcement learning algorithms to further optimize trading strategies. This will include the application of deep reinforcement learning and exploration of various reward structures to enhance strategy performance.

References

[1] Marco Avellaneda and Jeong-Hyun Lee. Statistical arbitrage in the us equities market. *Quantitative Finance*, 10(7):761–782, 2010.

[2] Victor Chang, Xiaowen Man, Qianwen Xu, and Ching-Hsien Hsu. Pairs trading on different portfolios based on machine learning. *Expert Systems*, 38(3):e12649, 2021.

[3] Binh Do and Robert Faff. Are pairs trading profits robust to trading costs? *Journal of Financial Research*, 35(2):261–287, 2012.

[4] Binh Do, Robert Faff, and Kais Hamza. A new approach to modeling and estimation for pairs trading. In *Proceedings of 2006 financial management association European conference*, volume 1, pages 87–99. Citeseer, 2006.

[5] R.J. Elliott, J. Van Der Hoek, and W.P. Malcolm. Pairs trading. *Quantitative Finance*, 5(3): 271–276, 2005.

[6] Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica*, 55(2):251–276, 1987.

[7] Eugene Fink and Harith Suman Gandhi. Important extrema of time series. 2007.

[8] Alexander Galenko, Elmira Popova, and Ivilina Popova. Trading in the presence of cointegration. *The Journal of Alternative Investments*, 15(1):85–97, 2012.

[9] Evan Gatev, William N Goetzmann, and K Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827, 2006.

[10] Jorge Guijarro-Ordonez, Markus Pelger, and Greg Zanotti. Deep learning statistical arbitrage. *Available at SSRN 3862004*, 2021.

[11] Nicolas Huck and Komivi Afawubo. Pairs trading and selection methods: is cointegration superior? *Applied Economics*, 47(6):599–613, 2015.

[12] T. Kanamura, S.T. Rachev, and F.J. Fabozzi. A profit model for spread trading with an application to energy futures. *The Journal of Trading*, 5(1):48–62, 2010.

[13] Tim Leung and Xin Li. Optimal mean reversion trading with transaction costs and stop-loss exit. *International Journal of Theoretical and Applied Finance*, 18(03):1550020, 2015.

[14] Tim Leung and Xin Li. *Optimal Mean Reversion Trading: Mathematical Analysis and Practical Applications*. Modern Trends in Financial Engineering. World Scientific Publishing Company, 2016.

[15] Tim Leung and Hung Nguyen. Constructing cointegrated cryptocurrency portfolios for statistical arbitrage. *Studies in Economics and Finance*, 2019.

[16] Rong Qi Liew and Yuan Wu. Pairs trading: A copula approach. *Journal of Derivatives & Hedge Funds*, 19(1):12–30, 2013.

[17] G. Montana and K. Triantafyllopoulos. Dynamic modeling of mean reverting spreads for statistical arbitrage. *Computational Management Science*, 8:23–49, 2011.

[18] Simão Moraes Sarmento and Nuno Horta. Enhancing a pairs trading strategy with the application of machine learning. *Expert Systems with Applications*, 158:113490, 2020.

[19] Ganapathy Vidyamurthy. *Pairs Trading: quantitative methods and analysis*, volume 217. John Wiley & Sons, 2004.

[20] Wenjun Xie, Rong Qi Liew, Yuan Wu, and Xi Zou. Pairs trading with copulas. *The Journal of Trading*, 11(3):41–52, 2016.

[21] Jize Zhang, Tim Leung, and Aleksandr Aravkin. Sparse mean-reverting portfolios via penalized likelihood optimization. *Automatica*, 111:108651, 2020.