HIGH-FREQUENCY STATISTICAL ARBITRAGE

PROJECT REPORT

Boyu Yang
The Chinese University of Hong Kong, Shenzhen
boyuyang@link.cuhk.edu.cn

December 29, 2022

ABSTRACT

Statistical arbitrage is a long-horizon trading methodology that generates a low-risk profit and is designed to exploit persistent market anomalies. From a high-frequency perspective, calculations are often required to be performed in real time. Numerical methods such as simulation or quadrature post high demand on the computation speed within a short time constraint, increasing the difficulty of implementing statistical arbitrage. In this work, we explore beyond traditional pairs trading in the context of intraday data. We prove the existence of statistical arbitrage and derive optimality conditions with fixed transaction costs for trading a mean-reverting Ornstein Uhlenbeck process. Backtesting results on fifteen similar stocks' intraday data show that statistical arbitrage opportunities still exist in the high-frequency area with favorable profit and loss (PnL) after considering real transaction costs including market impact on the limit order book, commission fees, and bid-ask spread. Empirical results also show that the optimal analytical strategy outperforms the benchmark of traditional pairs trading, namely, the paradigm approach that uses asymmetric bands for trading a mean reverting process. The code and data are available at https://github.com/bradleyboyuyang/Statistical-Arbitrage.

Keywords Statistical Arbitrage · Pairs Trading · High-frequency Trading · Ornstein-Uhlenbeck Process · Mean Reverting Process · First-Passage Time · Profit and Loss (PnL)

1 Introduction

Compared with the traditional trend-following trading strategy such as momentum and reversal trading, arbitrage has long been a renowned quantitative trading strategy that usually gains relatively secure profit with little drawdown. Statistical arbitrage is one of the pillars of arbitrage trading, and has long been used by hedge funds and investment banks. The term statistical arbitrage encompasses a wide variety of investment strategies, which aims to identify and exploit temporal price differences between similar assets through statistical models.

The simplest form of statistical arbitrage is known as "pairs trading". Two stocks are selected that are "similar", usually based on historical co-movement in their price series. When the spread

between their prices widens, the arbitrageur sells the winner and buys the loser. If their prices move back together, the arbitrageur will profit. However, while Wall Street has developed a plethora of proprietary tools for sophisticated arbitrage trading, there is still a lack of understanding of how much arbitrage opportunity is actually left in financial markets. Besides, in recent years, there has been a growing number of high-quality, high-frequency datasets on high-frequency trading (HFT). In particular, the limit order book (LOB), a system of limit order status at different prices, which contains the most microscopic trading information provided by market participants. Hence, whether more complex and delicate statistical arbitrage models can still guarantee a favorable PnL remains obscure, especially in high-frequency areas.

This work focuses on the employment of statistical arbitrage on intraday minute stock data with a limit order book structure. We start from the traditional pairs trading with the cointegration test and time series analysis. Then we forward to derive the analytical formulae for optimal statistical arbitrage based on continuous mean-reverting models. The empirical backtesting results on intraday stock data will be compared and displayed near the end. Moreover, we include a fixed transaction cost in our model derivation to improve robustness. In strategy backtesting, we take into account various transaction costs including market impact, trade fees, and bid-ask spread. Compared with previous work, where the arbitrage models are either too delayed to capture price differences or too ideal regardless of market frictions, we strive for a balance between optimal analytical solutions and real implementations to make the model usable in real-world trading.

2 Literature Review

Low-frequency Statistical Arbitrage Traditional pairs trading is first developed by scientists at Morgan Stanley in the mid-1980s. Following Gatev et al. [7, 2006], the underlying mechanism is based on a two-stage procedure. First, find pairs of synchronous stocks whose prices have historically moved together. Second, observe the price spreads, i.e., the difference of normalized prices, in the following out-of-sample trading period. The strategy takes long in the undervalued stock and takes short in the overvalued stock. As long as the history repeats itself, the price spread will revert to its historical equilibrium and a profit is made. The strategy is now adapted in several ways afterward to compete with other market participants in the modern financial market.

Krauss [10, 2017] categorizes upgraded model form in the following approaches: Distance method, cointegration, time series, stochastic control, and others. Key contributions to these upgraded models come from Gatev et al. [7, 2006], Vidyamurthy [16, 2004], Elliott et al. [5, 2005], Avellaneda and Lee [1, 2010], Do and Faff [3, 4, 2010, 2012], and Pizzutilo [12, 2013] - all of them focus on daily data. Medium to low-frequency models usually showcase relative robustness regarding transaction costs. Models do not iterate and update frequently and the strategy return is usually stable.

High-frequency Statistical Arbitrage In recent years, more sophisticated techniques for analyzing data and exponential increase in computing power allow high-frequency trading. Thus, more and more scholars are shifting their focus to build high-frequency models that capture temporal price differences and market imperfections. In particular, William [2, 2010] lays a solid theoretical foundation that gives the optimal entry and exit points for Ornstein Uhlenbeck processes. Göncü et al. [8, 2016] extends the results of William to trade the spread portfolio of two assets. Furthermore, they measure the asymptotic probability of loss and conclude that uncertainty in the long-term mean and spread volatility restrains arbitrage opportunities. Johannes and Jens [14, 2017] applies different traditional strategies to minute-by-minute prices of S&P 500 constituents from 1998 to

2015 and achieves an annualized Sharpe ratio of 8.14 after transaction costs. In recent years, more researchers move to limit order book data and utilize relevant time series behavior. Wang et al. [17, 2021] design a statistical arbitrage strategy based on an improved version of Order Flow Imbalance (OFI) signal.

Compared with low-frequency statistical arbitrage, high-frequency models are more likely to discover intraday trading opportunities. Yet, these models often suffer from two extremes in real-world applications. First, most models in the literature are too ideal to be implemented in high-frequency areas in terms of either over simplicity (Gaussian assumptions, zero slippage, etc.) or over complexity (extensive computations, numerical simulations, etc.). Second, high-frequency models are much more sensitive to transaction costs such as trade fees and bid-ask spread. The model needs to be iterated and adapted to the changing hyperparameters regularly.

Models Based on Machine and Deep Learning With the development of large-scale AI models, another group of researchers focuses on employing machine and deep learning techniques to extract time series signals from price differences. Avellaneda and Lee [1, 2010] generate trading signals using Principle Component Analysis (PCA) and regressing stock returns on sector Exchange Traded Funds (ETFs) in US equities and then model the idiosyncratic returns as mean-reverting processes. Krauss et al. [11, 2017] takes a more aggressive approach by constructing long-short portfolios based on predicted cross-sectional return using random forests, gradient-boosted trees, neural networks, and ensemble models. Later, similar methods are used again by Fischer et al. [6, 2019] in cryptocurrency markets. More complicated models have been used in recent years, such as the convolutional transformer used by Jorge et al. [9, 2021] to extract time series signals from residual portfolios that are constructed from multi-factor models.

3 Proposed Model

In this section, we first derive the traditional pairs trading model with cointegration tests and time series analysis. Then we derive the analytical solution for optimal statistical arbitrage based on continuous Ornstein Uhlenbeck models.

3.1 Spread Models

Spread models are the most widely used classical models that trade the spread portfolio under cointegration tests and time series analysis. Coingration tests are first applied to find similar assets to construct the spread portfolio. Then the spread portfolio is traded when its market value goes beyond or below a certain threshold. In most simplified cases, the thresholds are set as a constant amount of standard deviation, e.g., two times the standard deviation from the mean.

Cointegration Coingration tests identify scenarios when two or more non-stationary time series are integrated together in a way that they cannot deviate from equilibrium in the long term. A typical test for cointegration usually is the Engle-Granger two-step method. In Avellaneda and Lee [1, 2010], the cointegration between two stocks is defined as

$$ln(S_t^A/S_0^A) = \alpha(t - t_0) + \gamma ln(S_t^B/S_0^B) + \epsilon_t \tag{1}$$

which is indeed a linear regression of the log-return of asset A towards that of asset B. The excess return of the spread position is therefore

$$X_t = ln(S_t^A/S_0^A) - \gamma ln(S_t^B/S_0^B) - r_f t \tag{2}$$

where r_f is the riskfree rate. Concerning the continuous time horizon we consider, the riskfree rate r_f is negligible. Further drop the constant term involving the initial prices, we simply use $X_t = ln(S_t^A) - ln(S_t^B)$ for generating the buy and sell signals.

Neutral Positions When the value of spread portfolio goes beyond the desired threshold, we expect its value to fall into a regular interval near the mean. Hence we will take long positions in the undervalued stock and short positions in the overvalued stock accordingly. Yet, the amount traded should obey the "definition" of an arbitrage strategy, which usually requires the overall positions to be "neutral". Denote the hedge ratio H as the ratio of positions between asset A to asset B. For simplicity, we do not consider taking leverage, namely, borrowing or lending. Then the following expression for H describes two typical cases that suit the "neutrality", which are called cash neutral and market/delta neutral respectively,

$$H = \begin{cases} S_A(t^*)/S_B(t^*) \\ \beta_A/\beta_B \end{cases} \tag{3}$$

where t^* is any trading point and β is the market factor exposure. In the upcoming model validation, we take the leverage ratio to be cash neutral, which means we are matching the size of long and short positions so that the cash flows cancel out. Under the additional guarantee of cointegration tests, the pairs trading strategy is market neutral in nature, indicating that the direction of the overall market does not affect the PnL. Each time we open a cash-neutral position to obtain a nearly zero initial cost of setting up the arbitrage portfolio (possibly larger than zero because of transaction costs). Then the PnL can be easily calculated based on the entry and exit points.

3.2 Ornstein-Uhlenbeck Process

Basic Model In the upcoming derivation, we assume a stationary mean-reverting process (such as logarithmic underlying asset price, price differences, residual portfolio, etc.). Ornstein-Uhlenbeck (OU) process, also known as the Vasicek model, is a classical stochastic model that denlineates the mean-reverting behavior with a differential term and a Wiener process term. The long-term mean and a constant model volatility are captured in model parameters. We first briefly derive the analytical solutions for OU process. Suppose a stochastic process X_t satisfies the following differential equation

$$dX_t = \kappa(\theta - X_t)dt + \sigma dW_t \tag{4}$$

Then by Itô's Lemma,

$$X_t = X_0 e^{-\kappa t} + \theta(1 - e^{-\kappa t}) + \sigma \int_0^t e^{\kappa(u-t)} dW_u \tag{5}$$

Since the remaining intergral is an Itô integral which follows a normal distribution indicated by Itô's Lemma, we have that r_t also follows a normal distribution

$$X_t \sim N\left(X_0 e^{-\kappa t} + \theta(1 - e^{-\kappa t}), \sigma^2 \frac{1 - e^{-2\kappa t}}{2\kappa}\right) \tag{6}$$

The take-away message is that Ornstein-Uhlenbeck process models the mean-reverting behavior with a Gaussian structure. The model requires a stable long-term mean and volatility to guarantee stability and robustness of model parameters.

First-Passage Time of Ornstein-Uhlenbeck Process The first-passage time of a stochastic process has been widely studied. For a process started at $x = c$ with a barrier at $x = a$, the first-passage time is defined as follows,

$$T_{a,c} = \inf\{t \geq 0 : X_t = a | X_0 = c\}$$

It refers to the first time that a process exceeds a certain level with a given initial state. For insurance companies, the estimation of ruin probability is related to the first-passage time that the net loss ever surpasses initial reserve. For standard Wiener process, we care about the first hitting time T_a when the process $B(t)$ reaches level a. For OU process in equation 4, we first normalize the model to a nondimentional state, that is, $Z_t = (X_t - \theta)/(\sigma/\sqrt{2\kappa})$ and $\tau = \kappa t$. Then the standardized OU prcess is given as

$$dZ_\tau = -Z_\tau d\tau + \sqrt{2} dW_\tau \tag{7}$$

we use the first passage time density and Laplace transform given in Finch (2014) for the standardized OU process. Let $f_{a,c}(t)$ denote the density function for $T_{a,c}$. The general explicit form for $f_{a,c}(t)$ is usually difficult to express. In the special case when $a = 0$, namely, the dimensionless OU process firstly goes to its mean zero, the probability density of the scaled first passage time $T_{0,c}$ starting from c is given by

$$f_{0,c} = \sqrt{\frac{2}{\pi}} \frac{|c| e^{-t}}{(1 - e^{-2t})^{3/2}} \exp\left(-\frac{c^2 e^{-2t}}{2(1 - e^{-2t})}\right)$$

However, for $a \neq 0$, formulae for $f_{a,c}(t)$ is difficult to express explicitly. Thomas [15, 1975], Ricciardi and Sato [13, 1988] derive the analytical solution for the mean and variance when $a > 0$ and $c > 0$, given in the following equations where erf(x) is the error function.

$$\mathbb{E}[T_{a,0}] = \sqrt{\frac{\pi}{2}} \int_0^a \left(1 + \text{erf}\left(\frac{t}{\sqrt{2}}\right)\right) \exp\left(\frac{t^2}{2}\right) dt = \frac{1}{2} \sum_{k=1}^{\infty} \frac{\left(\sqrt{2}a\right)^k}{k!} \Gamma\left(\frac{k}{2}\right) \tag{8}$$

$$\mathbb{E}[T_{0,c}] = \sqrt{\frac{\pi}{2}} \int_{-c}^0 \left(1 + \text{erf}\left(\frac{t}{\sqrt{2}}\right)\right) \exp\left(\frac{t^2}{2}\right) dt = \frac{1}{2} \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\left(\sqrt{2}c\right)^k}{k!} \Gamma\left(\frac{k}{2}\right) \tag{9}$$

and

$$\mathbb{V}[T_{a,0}] = \mathbb{E}[T_{a,0}]^2 - \frac{1}{2} \sum_{k=1}^{\infty} \frac{\left(\sqrt{2}a\right)^k}{k!} \Gamma\left(\frac{k}{2}\right) \Psi\left(\frac{k}{2}\right) \tag{10}$$

$$\mathbb{V}[T_{0,c}] = \frac{1}{2} \sum_{k=1}^{\infty} (-1)^k \frac{\left(\sqrt{2}c\right)^k}{k!} \Gamma\left(\frac{k}{2}\right) \Psi\left(\frac{k}{2}\right) - \mathbb{E}[T_{0,c}]^2 \tag{11}$$

Here $\Psi(x) = \psi(x) - \psi(1) = \Gamma(z)'/\Gamma(z) - \Gamma(1)'/\Gamma(1)$ and $\psi(x)$ is the digamma function. In particular, $\Psi(1) = 0$ and

$$\Psi(x) = \begin{cases} \sum_{k=1}^{x-1} \frac{1}{k}, & x \text{ is an integer} > 1 \\ -2ln(2) + 2 \sum_{k=1}^{x-1/2} \frac{1}{2k-1}, & x \text{ is a half-integer} > 0 \end{cases}$$

3.3 Stochastic Models

After discussing the traditional pairs trading spread models, we now derive the optimality statistical arbitrage solution based on the previously introduced Ornstein-Uhlenbeck process.

3.3.1 Notation

We first list the notations that will be used through the model derivation.

Attribute	Description
X_t	logarithmic price of the underlying
α, η	model parameters for the Vasicek model
a, m	entry and exit point for an arbitrage strategy
T	total trade length, time taken to complete a trade cycle
c	fixed transaction costs
$r(a, m, c)$	return per trade, a function of trade point and cost
$\mu(a, m, c)$	expected return for the arbitrage strategy
$\sigma(a, m, c)$	variance for the arbitrage strategy

Table 1: Preliminary notations for continuous time trading model

3.3.2 Continuous Time Trading Model

We now propose the theoretical optimal statistical arbitrage trading. Assume the logarithmic price of the traded security p_t satisties the Vasicek model with the long-term drift term $\theta = 0$, namely,

$$log(p_t) = X_t; \qquad X_{t_0} = x_0 \tag{12}$$

$$dX_t = -\alpha X_t dt + \eta dW_t \tag{13}$$

where $\alpha > 0, \eta > 0$ and W_t is the Wiener process. We now consider a continuous time statistical arbitrage strategy that enters a trade when $X_t = a$ and exits when $X_t = m$. Then at the next time when $X_t = a$ we complete a whole trading cycle. Since the Vasicek model describes a stationary process, such a strategy can be thought as periodic, since the arbitrage actions are repeated between trade entry points. An illustration graph is given in figure 1.

Without loss of generality, we assume that $a < m$, it is easy to conclude from figure 1 that the return per trade is a function of the entry price, exit price, and transaction cost, namely, $r(a, m, c) = (m - a - c)$. Since T gives the total time length of the trade cycle, $1/T$ gives the average number of trades per unit time. Therefore, the trade cycle length T is a renewal process which satisties the big-O notation

$$\mathbb{E}[1/T] \sim O\left(1/\mathbb{E}[T]\right)$$

$$\mathbb{V}[1/T] \sim O\left(\mathbb{V}[T]/\mathbb{E}[T]^3\right)$$

Hence, the expected return and variance of return per unit time are given by

$$\mu(a, m, c) = r(a, m, c)\mathbb{E}[1/T] = r(a, m, c)/\mathbb{E}[T] \tag{14}$$

$$\sigma^2(a, m, c) = r(a, m, c)^2\mathbb{V}[1/T] = r(a, m, c)\mathbb{V}[T]/\mathbb{E}[T]^3 \tag{15}$$

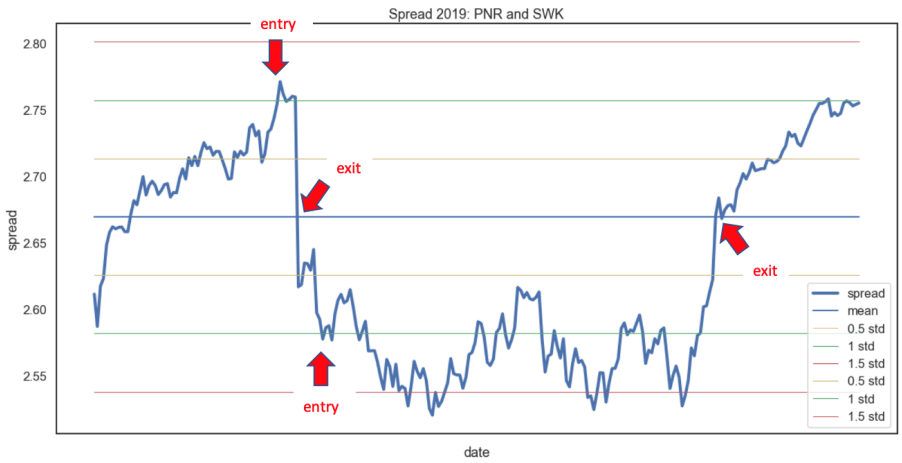

Figure 1: Illustration of the entry and exit point with respect to a stationary process

Decompose the total trade cycle T into T_1, T_2 as the time from a to m and the time from m back to a respectively. Since X_t is Markovian, the passage time T_1, T_2 are independent because of non-overlapping time intervals.

$$\mathbb{E}[T] = \mathbb{E}[T_1] + \mathbb{E}[T_2] \tag{16}$$
$$\mathbb{V}[T] = \mathbb{V}[T_1] + \mathbb{V}[T_2] \tag{17}$$

From the definition of first-passage time, T_1, T_2 can be identified as first-passage times for the process X_t with respective starting points and barrier points to be (a, m) and (m, a) respectively. We can first normalize the original Vasicek model 13 to a nondimentional process as equation 7 does

$$dY_\tau = -Y_\tau d_\tau + \sqrt{2} dW_\tau$$

Here, the scaled process $Y_t = X_t\sqrt{2\alpha}/\eta$, time dilation $\tau = \alpha t$, scaled entry level $\bar{a} = a\sqrt{2\alpha}/\eta$, exit level $\bar{m} = m\sqrt{2\alpha}/\eta$, transaction cost $\bar{c} = c\sqrt{2\alpha}/\eta$, and trade length $\bar{T} = \alpha T$. Assume the first-passage time $T_{a,m} = \inf\{t \geq 0 : Y_t > a | Y_0 = m\}$. Using the first-passage time formulae from equation 8 to equation 11, together with the summation property by 16 and 17,

$$\mathbb{E}[T_{a,m}] = \begin{cases} \phi_1(a) - \phi_1(m); & a > m \\ \phi_1(-a) - \phi_1(-m); & a < m \end{cases} \tag{18}$$

$$\mathbb{V}[T_{a,m}] = \begin{cases} \phi_1(a)^2 - \phi_2(a) + \phi_2(m) - \phi_1(m)^2; & m < a \\ \phi_1(-a)^2 - \phi_2(-a) + \phi_2(-m) - \phi_1(-m)^2; & m > a \end{cases} \tag{19}$$

where $\phi_1(x)$ and $\phi_2(x)$ are given by,

$$\phi_1(z) = \frac{1}{2}\sum_{k=1}^{\infty} \Gamma\left(k/2\right) \left(\sqrt{2}z\right)^k / k!$$
$$\phi_2(z) = \frac{1}{2}\sum_{k=1}^{\infty} \Gamma\left(k/2\right) \Psi\left(k/2\right) \left(\sqrt{2}z\right)^k / k!$$

Therefore we can calculate the expected trade length

$$\mathbb{E}[\bar{T}] = \pi \left(\text{Erfi}\left(\bar{m}/\sqrt{2}\right) - \text{Erfi}\left(\bar{a}/\sqrt{2}\right)\right)$$

Scale the solution back to the original model, we get the closed form solutions of the trade length

$$\mathrm{E}[T] = \frac{\pi}{\alpha}\left(\text{Erfi}\left(m\sqrt{\alpha}/\eta\right) - \text{Erfi}\left(a\sqrt{\alpha}/\eta\right)\right) \tag{20}$$

Substitute back to the return equation 14 and risk equation 15, we have the expected return for the strategy

$$\mu(a,m,c) = \frac{\alpha(a-m-c)}{\pi\left(\text{Erfi}\left(m\sqrt{\alpha}/\eta\right) - \text{Erfi}\left(a\sqrt{\alpha}/\eta\right)\right)} \tag{21}$$

The explicit expressions for $\sigma(a,m,c)^2$ and the strategy variance can be calculated similarly and are omitted here due to limited space.

3.3.3 Optimal Strategy

In this section, we analyze in detail the optimal statistical arbitrage strategy that maximizes expected return $\mu(a,m,c)$. Based on the results in equation 21, the optimization problem is as follows

$$\max_{a,m} \frac{\alpha(a-m-c)}{\pi\left(\text{Erfi}\left(m\sqrt{\alpha}/\eta\right) - \text{Erfi}\left(a\sqrt{\alpha}/\eta\right)\right)} \text{ subject to } a<0, m>0 \tag{22}$$

The first order condition gives

$$\sqrt{\frac{4\pi}{\alpha\eta^2}} e^{\frac{\alpha a^2}{\eta^2}}(m-a-c) - \frac{\pi}{a}\left(\text{Erfi}\left(\frac{m\sqrt{\alpha}}{\eta}\right) - \text{Erfi}\left(\frac{a\sqrt{\alpha}}{\eta}\right)\right) = 0$$
$$\sqrt{\frac{4\pi}{\alpha\eta^2}} e^{\frac{\alpha m^2}{\eta^2}}(m-a-c) - \frac{\pi}{a}\left(\text{Erfi}\left(\frac{m\sqrt{\alpha}}{\eta}\right) - \text{Erfi}\left(\frac{a\sqrt{\alpha}}{\eta}\right)\right) = 0$$

We have $m=-a$ is the solution. Thus the optimal entry and exit bands are symmetric about zero, which is counter-intuitive as traditional paradigm for pairs trading uses asymmetric bands, entering a trade when the process exhibits a threshold and exiting when it returns to zero. Therefore, the optimal entry point a satisfies

$$e^{\frac{\alpha a^2}{\eta^2}}(2a+c) = \eta\sqrt{\frac{\pi}{\alpha}}\left(\text{Erfi}\left(\frac{a\sqrt{\alpha}}{\eta}\right)\right) \tag{23}$$

While the root of the equation can be solved numerically, an approximate solution can be obtained through a simple 3^{rd} order Taylor expansion at $a=0$,

$$c + \frac{\alpha c}{\eta^2}a^2 + \frac{4\alpha}{3\eta^2}a^3 = 0$$

Besides, the analytical solution when the objective function is the Sharpe ratio can be deducted similarly, as illustrated by William [2, 2010]. The optimal entry and exit band are shown to be symmetric again in this case, yet the optimal a is less tractable and the first order condition can only be solved numerically. Other more complicated objective function such as conditional value at risk (CVaR) can be applied, but the analytical solution may not be so forthcoming.

4 Model Validation

4.1 Data Source and Processing

Limit order book data are generally not publicly available. We use the five minutes snapshot of fifteen similar stocks (AA to ZZ) from a previous competition held by Optiver. We disclose summary statistics of the mid prices in table 4.1. First, we aggregate the data so that the snapshot is

	AA-price	BB-price	CC-price	AA-volume	BB-volume	CC-volume
count	16991	16991	16991	16991	16991	16991
mean	87.13	92.31	92.00	133.81	104.24	111.06
std	3.81	3.86	3.42	2.24	21.05	17.37
min	79.60	83.97	85.58	65.00	57.00	61.50
25%	83.20	89.50	89.47	119.50	93.00	98.50
50%	87.50	92.55	91.30	132.50	103.00	109.50
75%	91.00	95.35	94.67	146.50	114.50	122.50
max	93.12	100.90	99.20	248.00	171.00	187.00

Table 2: Summary statistics of mid price and volume for three stocks

continuous and available for every five minute timestamp in 24 hours. Then, we calculate the mid price for each stock. For each timestamp the data contain the best bid and ask price and volume during the five minutes period, the structure of which is identical to that of the limit order book. Hence, higher frequency data such as tick-level data can be analyzed similarly.

4.2 Model Implementation

4.2.1 Spread Models

Coingration Tests We first use the two-stage Engle-Granger test to select the desired trading pairs. First, we obtain the residual series of a time series regression with respect to the logarithmic stock mid prices. Then, we conduct the augement Dickey-Fuller (ADF) test to test for stationarity of the residuals. Since the null hypothesis is non-stationary residuals, we select the pairs with p-values less than 0.01. The corresponding regression coefficients and summary statistics of the residuals are given in table 4.2.1 where Gamma denotes the linear coefficient of the time series regression, Alpha denotes the constant of the residual regression, Crossing denotes number of times the residuals cross the long-term mean, Period denotes the reciprocal of Crossing, and Mean and Std give the long-run mean and volatility.

After cointegration tests, ten stocks pairs are left for further arbitrage backtesting. We draw the spread residual of HH and JJ as an example in figure 2, where the shaded blue area indicates the 95% confidence intervals.

Threshold Analysis Following the traditional pairs trading, the entry points are determined by a multiply of standard deviation with respect to the mean. The rule-of-thumb choice for the threshold is two times the standard deviation, yet in intraday data especially considering the commission fee, bid-ask spread, and market impact, the classical choice may differ. Therefore, we conduct a threshold analysis to determine the optimal threshold. A simple grid search algorithm is applied and the grids are set to be integer multiples of the standard deviation, as illustrated in figure 3.

Pairs	Constant	Gamma	Alpha	Crossing	Period	Mean	Std
(BB, DD)	-2.4182	1.4758	-0.0075	660	0.001515	-0.0	0.0100
(BB, JJ)	-2.4809	1.6206	-0.0000	1475	0.000678	0.0	0.0044
(DD, HH)	-1.6797	1.2787	-0.0040	396	0.002525	0.0	0.0097
(DD, JJ)	0.2277	1.0356	0.0002	592	0.001689	-0.0	0.0072
(FF, MM)	0.3061	0.9676	-0.0338	1413	0.000708	-0.0	0.0031
(FF, NN)	-2.6104	1.4881	-0.0022	705	0.001418	0.0	0.0061
(MM, NN)	-2.9034	1.5153	-0.0002	627	0.001595	-0.0	0.0068
(BB, HH)	-4.8181	1.8712	-0.0024	334	0.002994	-0.0	0.0181
(HH, JJ)	1.9492	0.7041	-0.0002	363	0.002755	0.0	0.0090
(AA, II)	0.0445	0.9944	-0.0002	363	0.002755	0.0	0.0214

Table 3: Summary statistics of residuals and p-values of ADF tests

Strategy Backtesting When designing the backtesting framework, we assume all the orders are market orders, that is, we pay the taker fee every transaction and the orders are filled at the opposite prices on the limit order book. Therefore the hedge ratio H mentioned in equation 3 is now

$$H = \gamma \times \left(\frac{S^A_{t,bid}}{S^B_{t,ask}} \right) \tag{24}$$

assuming we take a short position in stock A and a long position in stock B where γ is the regression coefficient in the cointegration tests mentioned in 1. To make the backtesting results comparable among different pairs of stocks with difference price magnitudes, we force that each time when we enter a long and short position, each side is worth around 10000. Hence the backtesting results

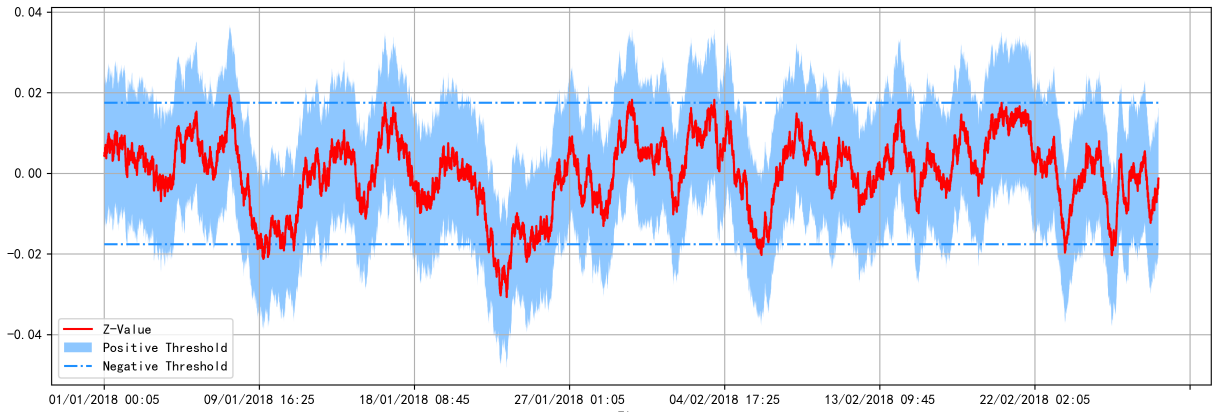

Figure 2: Residual time series of stock HH and JJ with 95% confidence intervals

become comparable in magnitude among different trading pairs and the overall PnL trend is not affected. Meanwhile, we take the bid and ask volume into consideration. Each time the exact trade amount is determined by both the worth limit 10000 and the actual bid and ask volume available.

4.2.2 Stochastic Models

For a standard OU-process 13, we adopt a simple ordinary least square method to estimate model parameters for simplicity. First, we approximate the stochastic differential equation by discretizing

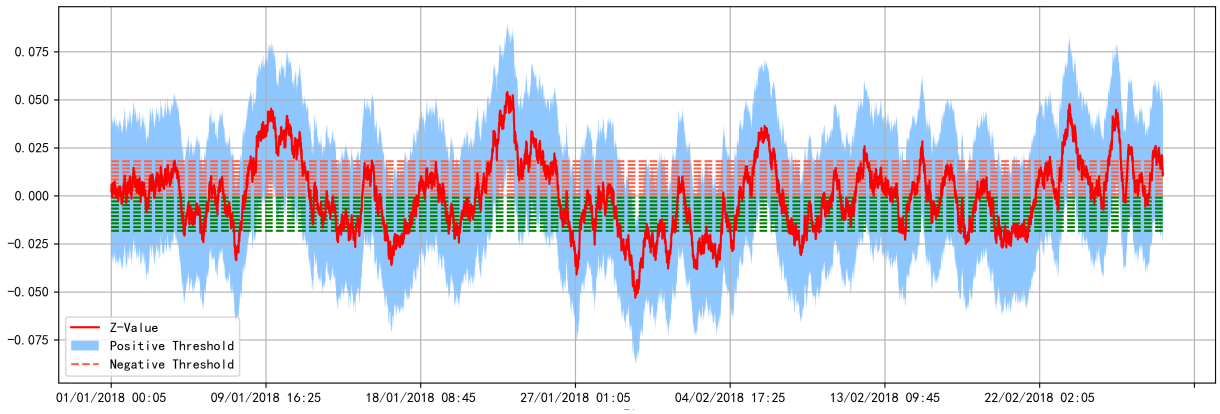

Figure 3: Thresholds with integer multiples of the standard deviation

the time grids, known as the Euler-Maruyama method,

$$\Delta X_t = \kappa\theta\Delta t - \theta X_t \Delta t + \sigma\epsilon_t$$

With fixed time time interval Δt, it is easy to see the regression specification where the regressor is X_t and the regressand is first-order difference series ΔX_t. The parameter estimation results for ten selected pairs are given in table 4.2.2. As illustrated in the optimal solutions of equation 23, the

Pairs	Constant	Gamma	Std	κ	θ	σ
(BB, DD)	-2.4182	1.4758	0.0100	0.00685	0.00008	0.00117
(BB, JJ)	-2.4809	1.6206	0.0044	0.03656	0.00001	0.00117
(DD, HH)	-1.6797	1.2787	0.0097	0.00342	-0.00000	0.00080
(DD, JJ)	0.2277	1.0356	0.0072	0.00598	-0.00002	0.00078
(FF, MM)	0.3061	0.9676	0.0031	0.03689	-0.00001	0.00083
(FF, NN)	-2.6104	1.4881	0.0061	0.00900	0.00001	0.00082
(MM, NN)	-2.9034	1.5153	0.0068	0.00701	0.00006	0.00081
(BB, HH)	-4.8181	1.8712	0.0181	0.00209	0.00025	0.00117
(HH, JJ)	1.9492	0.7041	0.0090	0.00212	-0.00016	0.00058
(AA, II)	0.0445	0.9944	0.0214	0.00154	0.00279	0.00108

Table 4: Ornstein-Uhlenbeck process parameter estimation results

optimal entry and exit band are symmetric. Hence, the backtesting logic is the same as the previous spread models, with the only difference in the entry and exit conditions. Similarly, we can either use the numerical solutions for 23 with the estimated parameters above or simply run a grid search for the optimal solution.

4.3 Results & Sensitivity Analysis

In this section, we show the backtesting results with bid-ask spread and commission fees and conduct a sensitivity analysis towards the total transaction costs. The grid of threshold is taken from 1 to 10 times of the standard deviation. The optimal thresholds and corresponding PnLs with the bid-ask spread considered are searched based on the overall return. We can further plot the change in positions as time evolves for each pair. Take the pair (DD, HH) as an example, the position

dynamics are shown in figure 4 under its optimal threshold. It can be seen that the strategy does not open positions frequently. We hold opposite positions most of the time and close positions sometimes when the spread portfolio falls back to the mean level, which is zero in our assumptions.

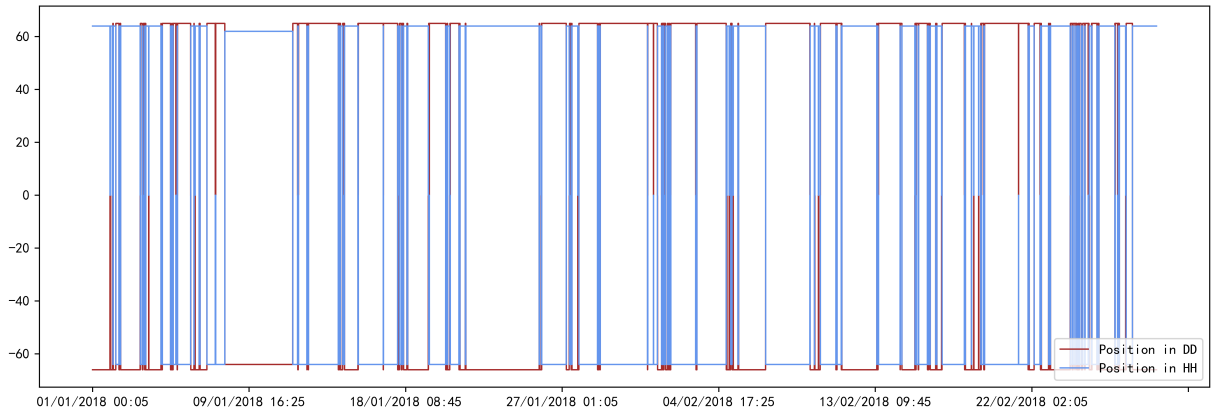

Figure 4: Position changes for stock DD and HH

Finally, we construct an equal-weighted portfolio of pairs. From the previous results of the ten pairs, we consider (BB, JJ), (FF, MM), (DD, HH), (AA, II) in our final portfolio to avoid that one stock appears in two or more pairs. Without considering commissions, the PnL curve and the total portfolio PnL are given in 5.

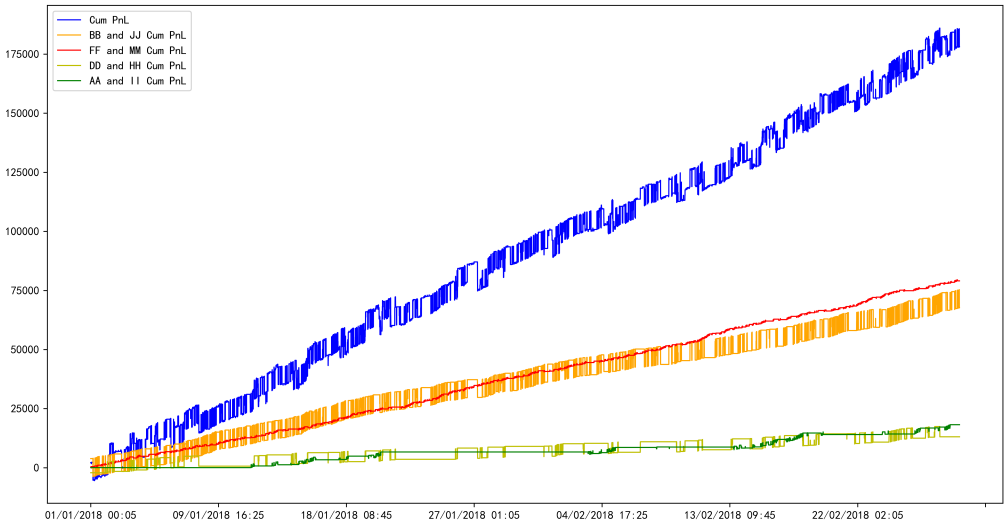

Figure 5: Overall PnL results with bid-ask spread

The cumulative portfolio PnL is 178207 in the two months period with a cash neutral and delta neutral initial investment in our statistical arbitrage trading. Then we add a fixed commission of 0.1% universally on all stocks. After considering a fixed commission rate together with the bid-ask spread, the PnL curve is shown in figure 6. Compared with figure 5, the portfolio PnL and that of each pair decreases. The PnL curve still follows similar trends, implying that the arbitrage model is robust enough. Yet, the equity curve displays larger volatility as the width of the PnL band increases. With an increasing transaction cost, the optimal threshold increases in order to guarantee that the strategy return can cover necessary transaction costs.

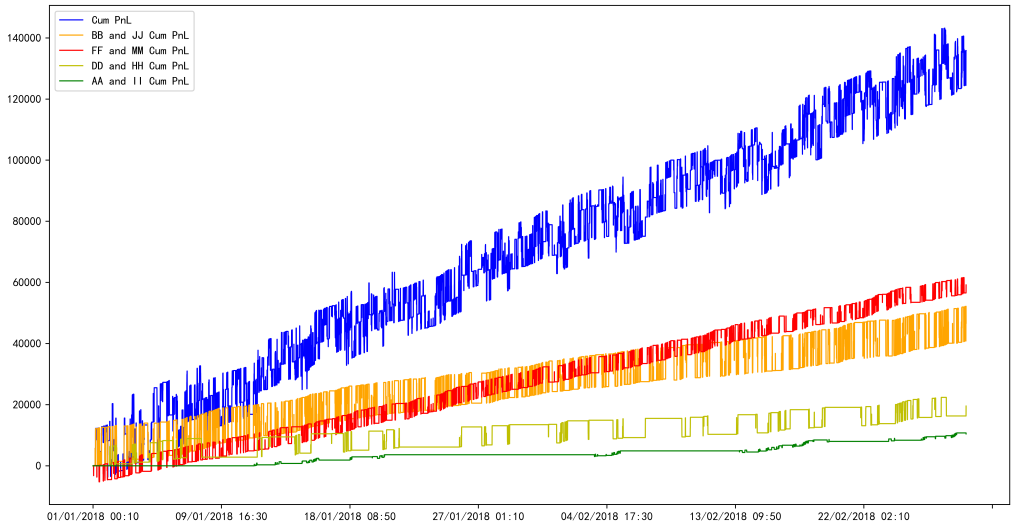

Figure 6: Overall PnL results with bid-ask spread and commission of 0.1%

4.4 Model Risks and Improvements

The largest risk of statistical arbitrage is when statistical models that are used to detect the "fair" price become ineffective in some extreme market conditions. We list several crucial risk exposure to consider when designing real-world statistical arbitrage strategies:

- The model's stochastic nature. For instance, although the expected return per trade is deterministic, the time frame over which the return is realized, namely, T, is stochastic. Hence a trade may take a long time before reaching the exit level and may significantly deviate from the desired exit level.
- Slippage. Here the term slippage incorporates the effect of bid-ask spread and market impact. The existence of slippage largely impairs the PnL of arbitrage in reality, resulting in severe competition for computation speed.
- Gaussian behavior. The Ornstein-Uhlenbeck model is Gaussian, yet the real spread portfolio is usually known as non-Gaussian.
- Discretization error. We are using continuous models to model discrete order book snapshots.

Here, we offer some possible ways to alleviate the various risk factors listed above. To decrease the discretization error, the data used are usually tick-level data. Tick data are more "continuous" but also face increasing microstructure noise problems. To deal with trading slippage, the calculation of the optimal entry and exit point can incorporate a fixed slippage directly. The fixed level usually depends on a company's overall ordering speed in the market. Besides, the optimal strategy derivation can also be applied to non-Gaussian processes such as the generalized Ornstein-Uhlenbeck model which is driven by a Levy noise. Yet the increased model accuracy is at the expense of increased computation requirements. In real high-frequency trading, numerical methods may not be fast enough to update calculations within the required time constraints. Hence, all the possible improvements mentioned need to be carefully measured with the corresponding trade-off.

5 Conclusion

In general, statistical arbitrage focuses on building statistical models that can measure the "fair" price and make a profit based on the discrepancy between the "fair" price and the real market price. Compared to risk-free arbitrage, statistical arbitrage takes a more aggressive way of securing arbitrage opportunities based on statistical models in exchange for greater exposure to model risks. We list the traditional spread models which use cointegration and time series analysis to construct arbitrage strategies, as well as the analytical formulae for the optimal continuous-time trading model. Empirical results on intraday stock data show the continuous time trading model with symmetric trading bands outperforms the spread model with asymmetric trading bands. Above all, arbitrageurs need to strive for a balance between optimal analytical solutions and real implementations to make their models usable in real-world trading.

References

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

[2] William K Bertram. Analytic solutions for optimal statistical arbitrage trading. *Physica A: Statistical mechanics and its applications*, 389(11):2234–2243, 2010.

[3] Binh Do and Robert Faff. Does simple pairs trading still work? *Financial Analysts Journal*, 66(4):83–95, 2010.

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

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

[6] Thomas Günter Fischer, Christopher Krauss, and Alexander Deinert. Statistical arbitrage in cryptocurrency markets. *Journal of Risk and Financial Management*, 12(1):31, 2019.

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

[8] Ahmet Göncü and Erdinç Akyıldırım. Statistical arbitrage with pairs trading. *International Review of Finance*, 16(2):307–319, 2016.

[9] Jorge Guijarro-Ordonez, Markus Pelger, and Greg Zanotti. Deep learning statistical arbitrage. *arXiv preprint arXiv:2106.04028*, 2021.

[10] Christopher Krauss. Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 31(2):513–545, 2017.

[11] Christopher Krauss, Xuan Anh Do, and Nicolas Huck. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, 259(2):689–702, 2017.

[12] Fabio Pizzutilo. A note on the effectiveness of pairs trading for individual investors. *International Journal of Economics and Financial Issues*, 3(3):763–771, 2013.

[13] Luigi M Ricciardi and Shunsuke Sato. First-passage-time density and moments of the ornstein-uhlenbeck process. *Journal of Applied Probability*, 25(1):43–57, 1988.

[14] Johannes Stübinger and Jens Bredthauer. Statistical arbitrage pairs trading with high-frequency data. *International Journal of Economics and Financial Issues*, 7(4):650–662, 2017.

[15] Marlin U Thomas. Some mean first-passage time approximations for the ornstein-uhlenbeck process. *Journal of Applied Probability*, 12(3):600–604, 1975.

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

[17] Qingxue Wang, Bin Teng, Qi Hao, and Yufeng Shi. High-frequency statistical arbitrage strategy based on stationarized order flow imbalance. *Procedia Computer Science*, 187:518–523, 2021.