Available Online: 11 November 2025 DOI: 10.54254/3049-5768/2025.29242

Optimized high-frequency quantitative pairs trading based on overnight jump detection in emerging markets

Zhao Zhang

Baiwang Jasmine Garden, Beijing, China

zhazhag00@gmail.com

Abstract. Under the background of increasing volatility in the financial market, the cross day jump poses a challenge to the threshold framework of the traditional paired trading strategy. This study distinguishes between intraday fluctuations and inter day jumps, uses Bayesian search to dynamically adjust the threshold, and combines sliding window and co-integration theory to build a medium and high frequency quantitative strategy, filling the gap in the research of medium and high frequency paired trading in emerging trading markets such as Kechuang 50. The empirical results in the 50 sector of science and technology innovation show that the strategy has both high-yield and risk control capabilities. After deducting the handling fees, the performance of the strategy is better than that of index investment, and the sharp rate has increased by 107%. However, the strategy is restricted by short selling of a shares, requires securities lending and has a high threshold, which is only applicable to large institutions; At the same time, there is a waste of handling fees caused by the overlapping of stock pairs. In the future, we can improve the unilateral strategy to adapt to market constraints, and introduce external data such as big language model and financial news to alleviate the lag of strategy adjustment.

Keywords: pairs trading, jump diffusion, Bayesian search, sliding window

1. Introduction: research background and significance

Under the background of increasing volatility in the financial market, the innovation of quantitative trading strategy has become the key to deal with the complex environment. As a typical market neutral strategy, paired trading realizes risk hedging by identifying the mean reversion characteristics of price linked assets, and its effectiveness has been verified in many markets around the world. The traditional strategy relies on a fixed threshold (such as Z-score ± 2) to generate trading signals, but the cross day jump of stock prices poses a serious challenge to this framework. Cross Day jump refers to the significant price fault between the closing price and the opening price of the adjacent trading days, which is manifested in the large deviation of the opening price of the next day from the previous closing price and no transaction filling range. It is more common in the A-share market due to the high proportion of retail investors, information asymmetry and other factors.

Yang and Malik, (2024) and Cerda et al. (2021) have shown that the classic paired trading strategy with a fixed threshold (gatev et al., 1999) has a significantly lower yield than the paired trading strategy with a dynamic threshold when the market switches between a bear market and a bull market. However, the selected markets for the backtesting of these two studies are the foreign exchange market and the cryptocurrency market, and the effectiveness of this strategy in the A-share market has not been proved. Therefore, this study hopes to study a trading strategy that can effectively cross the bear market and the bull market in the A-share market, and the most appropriate trading threshold can be selected when the A-share market is in different states.

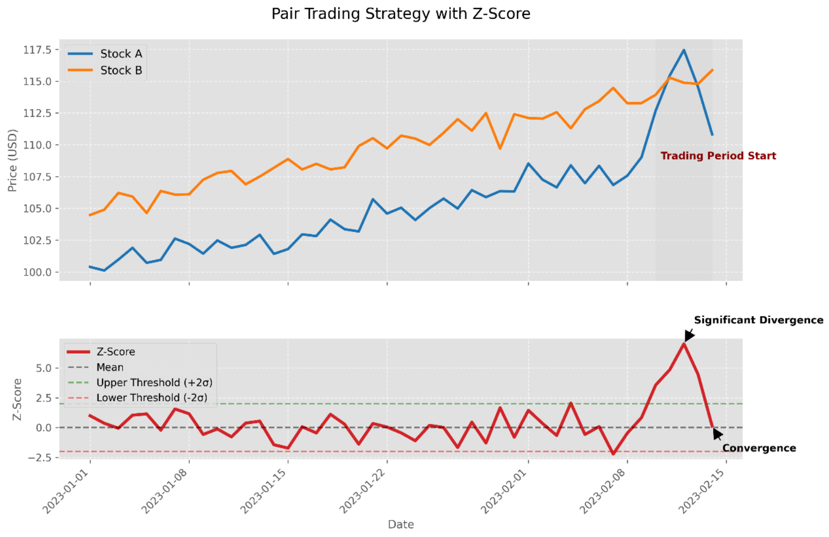
It is of great value to study the impact of the cross day jump of stock prices on the dynamic adjustment of thresholds: in theory, the jump detection technology and co-integration theory are integrated to break through the limitations of the static parameter framework; In practice, it helps investors identify jump risks and improve the robustness of strategies. Unlike previous studies, which do not distinguish between the types of volatility, this study distinguishes between intraday volatility and cross day jumps, and uses the trading threshold obtained by Bayesian search to replace the fixed trading threshold, making the strategy more consistent with the characteristics of China's A-share volatility that is greatly affected by policy news and frequent and violent style switching. But at the same time, it is worth noting that the strict restrictions on short selling in China's A-share market make it difficult to directly apply the traditional matching trading strategy. In the future, we can explore the dynamic adjustment mechanism of unilateral position based on co-integration relationship, so as to avoid short selling constraints and reduce the friction cost of bilateral transactions.

The innovation of this study is that it effectively combines Bayesian search and jump discovery technology, which makes it easier for this strategy to capture the changes of A-share market style, and adaptively adjust the trading threshold, so that the strategy can more effectively adapt to different A-share market styles.

2. Literature review

2.1. Research progress of pairing trading strategy

The evolution of paired transaction methodology can be divided into two stages: traditional static framework and modern dynamic optimization. The traditional method is represented by the co-integration test of Engle and Granger (1987) and the distance method of Gatev et al. (1999). By identifying price linked asset pairs, transactions are made when the price difference deviates from the threshold. However, the static threshold (such as ± 2σ) is difficult to adapt to the changes in market structure. After the financial crisis in 2008, the strategic sharp ratio fell by 40%, and the erosion of transaction costs further weakened profitability.



Z-Score = (Current Spread - Mean Spread) / Std Deviation of Spread | Lookback Period: 40 days | Trading Period: 5 days

Figure 1. Schematic diagram of paired transaction principle

This graph illustrates the basic logic of paired trading, where when there is a significant change in the price gap between two stock pairs with co-integration of price fluctuations, the gap tends to return to the mean after a period of time. At this point, going long on stocks that are relatively undervalued and going short on stocks that are relatively overvalued can be used for arbitrage.

2.2. Research on the detection and impact of stock price jump

Figueroa-López and Nisen (2020) proposed a new threshold kernel jump detection method based on the jump diffusion process. The method iteratively applies the threshold and kernel method in an approximate optimal way to improve the performance of finite samples. The impact of jump on trading strategy is significant: the jump factor of A-share market is negatively correlated with future earnings, and the cross-day jump makes the overnight fluctuation close to the intraday level. These findings provide a quantitative basis for the threshold adjustment of paired transactions, but existing studies have not systematically explored the relationship between cross day jumps and dynamic threshold optimization.

2.3. Current situation of crossover research

There are two gaps in the existing cross research: one is the lack of attention to the special impact of the cross-day jump. The literature focuses on intraday jumps and ignores the change of overnight intraday reversals on the spread structure. Second, there is no research on paired trading in the Chinese market. Eroğlu et al. (2023) chose the S&P 500 component stock in his research, and Dai et al. (2024) chose the constituent stocks of the Yuanta/P-shares Taiwan Top 50 ETF (0050.TW) and the Yuanta/P-shares Taiwan Mid-cap 100 ETF (0051.TW), the research market of Liu (2013) is soybean futures. A-shares have price limits and a relatively large proportion of retail investors are active in the market, so they have unique jumping characteristics, but lack of a targeted dynamic threshold model for paired trading.

2.4. Research issues and innovations

Core research question: how to effectively identify the cross day jump of stock price, and dynamically adjust the paired trading threshold based on it, so as to improve the robustness of A-share market strategy?

The innovation is reflected in the following aspects:

(1) for the first time, the power-law distribution characteristics of inter-day jumps are incorporated into the threshold model, breaking the normal distribution assumption;

(2) a multi-factor adjustment mechanism integrating jump intensity is proposed to replace the static threshold;

(3) the threshold search is optimized using Bayesian search, effectively solving the pain points of grid search requiring large computing resources when multiple parameters are used and random search being difficult to achieve global optimization;

(4) an end-to-end framework of "jump detection-threshold adjustment-strategy execution" is constructed, combining machine learning to improve real-time performance.

(5) the pair trading method is extended to the Science and Technology Innovation 50 sector, and an effective trial has been carried out in emerging markets such as A-shares.

3. Data and method

3.1. Data

This paper uses the minute level share price data of the constituent stocks of Kechuang 50 (from January 1, 2024 to January 1, 2025) and the dividend announcement of a shares, and restores the minute level share price to the previous compound share price data through the original data and dividend ex dividend announcement.

3.2. ADF inspection

The mathematical process of the augmented Dickey fuller test (ADF test) is to determine whether there is a unit root (i.e., non-stationarity) in the time series by constructing an autoregressive model, estimating key parameters and calculating test statistics. Different from ADF test, the original hypothesis sequence of KPSS test (Kwiatkowski-Phillips-Schmidt-Shin) is stable. When the sample size increases, ADF test has stronger ability to detect unit root, while KPSS test tends to reject the original hypothesis of stationarity. Since this paper uses minute level stock price data and the review period is 40 trading days, the sample size of the series is 40 * 4 * 60=9600. At this time, the effect of using ADF test will be better. Therefore, this paper chooses to use ADF test (see Appendix 1 for the detailed test process).

3.3. Correlation coefficient

There are two commonly used correlation coefficient calculations: Pearson correlation coefficient and Spearman correlation coefficient.

Spearman's rank correlation coefficient is a non-parametric statistic used to measure the monotonic relationship (not necessarily linear relationship) between two variables. Its mathematical process is based on the rank of the data rather than the original value. The core idea is to calculate the Pearson correlation coefficient after converting the variable to the rank.

Pearson correlation coefficient is a statistic that measures the strength and direction of the linear relationship between two continuous variables. Its mathematical process is based on the standardization of covariance and standard deviation.

The main difference between Pearson correlation coefficient and Spearman correlation coefficient is that Pearson correlation coefficient can better capture the linear relationship between the two sequences, while Spearman correlation coefficient can better capture the nonlinear relationship. Since the paired trading strategy in this paper is mean regression rather than trend trading, this strategy prefers to capture the linear relationship of stock price series, so the Pearson correlation coefficient is selected as the correlation test. The mathematical definition of Pearson correlation coefficient r is:

$$r = \frac{Cov(X,\ Y)}{\sigma_X \sigma_Y} \tag{1}$$

Cov (x, y) is the covariance of variables X and y, which reflects the degree of their collaborative changes;

σ_X and σ_Y are the standard deviations of X and y, respectively, used to eliminate dimensional effects and standardize covariance.

3.4. Jump discovery

The standard uses the jump threshold detection method to detect the jump of the overnight fluctuation of stocks. The essence of the threshold jump detection method is that as long as the absolute value of the process increment exceeds the threshold, the jump will occur. It can be used as an approximate and simple method to estimate the jump threshold according to the average value and variance of the volatility (Figueroa-López and Nisen, 2020). Since the review period in this paper only has 40 trading days and the amount of data is small, the 0.95 quantile of the cross-day fluctuation range in the review period is used as the jump threshold.

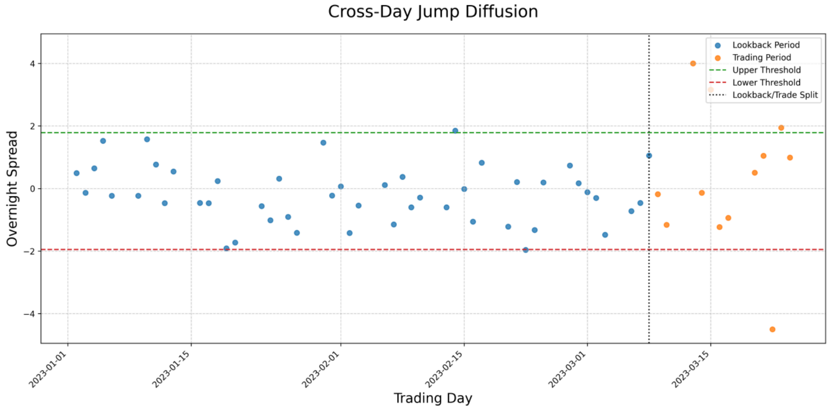


Figure 2. Overnight jump diffusion diagram

This figure shows the principle of overnight jump detection. The stock price fluctuations during the lookback period are used as a benchmark to find the jump threshold, which is then used as the jump detection threshold during the trading period.

3.5. EWMA

EWMA (exponentially weighted moving average) is a method for smoothing time series data. Its core idea is to give higher weight to recent data, and the weight of historical data decays exponentially with time. Because this paper filters the minute share price during the trading period, and there are up to 100 stock pairs in the portfolio, it is very sensitive to the amount of calculation of the filtering noise reduction algorithm during the trading period. Compared with other commonly used noise reduction algorithms, EWMA has a very small amount of calculation, which can well meet the requirements of low delay under high-frequency trading. Therefore, the modified algorithm is selected as the noise reduction algorithm (refer to Appendix 3 for detailed calculation and derivation formula).

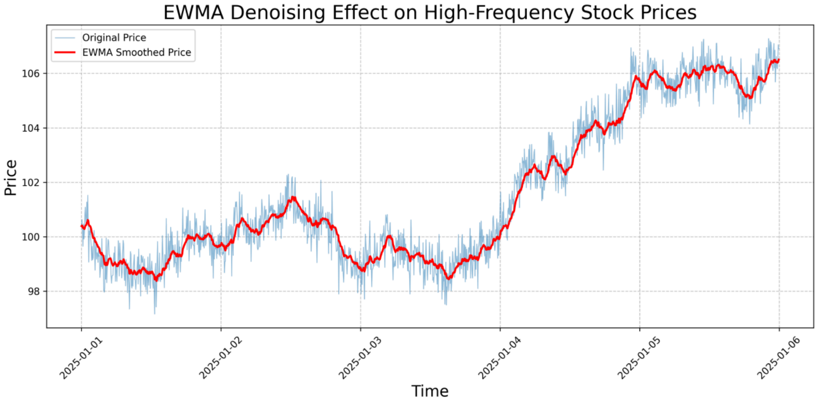


Figure 3. Schematic diagram of EWMA noise reduction for stock price fluctuation

This graph demonstrates the denoising effect of EWMA, which effectively filters out high-frequency noise in stock price fluctuations while preserving low-frequency stock price trends.

3.6. Bayesian search

Bayesian optimization is a sequential optimization method based on Bayesian theorem. By constructing the probability model of the objective function, it can efficiently explore the parameter space to find the global optimal solution. The advantage of Bayesian search is that the optimized function f(x) can be optimized as a black box without understanding the internal operation rules. At the same time, compared with grid search, when there are many parameters, Bayesian search can greatly reduce the consumption of computing resources. Compared with random search, Bayesian search can better search for the global optimal parameters (refer to Appendix 4 for the detailed search process).

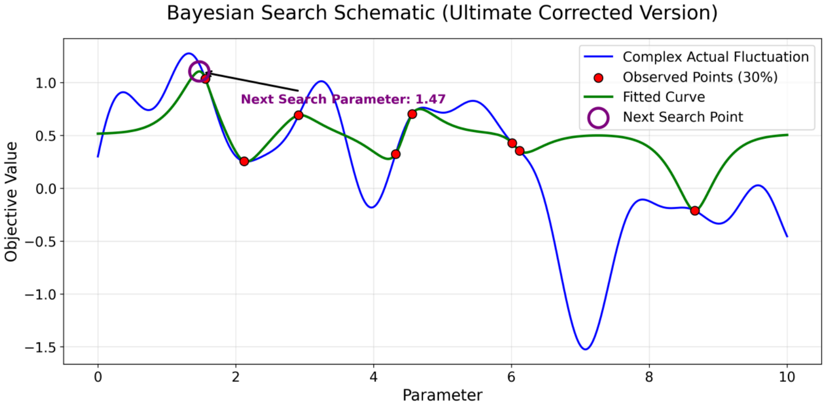


Figure 4. Bayesian search diagram

This figure shows the principle of Bayesian search. The proxy function is drawn according to the sampling point, and the parameter of the maximum value of the proxy function is used as the next sampling point, and the proxy function is continuously updated.

4. Pairing method

4.1. Method overview

As one of the core strategies of statistical arbitrage, the core of paired trading is to identify stock pairs with highly correlated price behavior, and use the opportunity of short-term deviation from the historical equilibrium relationship to trade. The pairing strategy framework proposed in this study integrates traditional statistical methods and modern dependency structure modeling technology, and constructs a robust trading portfolio through a three-stage screening process: first, the sliding window and correlation analysis are used to preliminarily screen candidate stock pairs, then copula theory is introduced to capture the nonlinear dependency structure, and finally the unit root test is combined to ensure the stability of the price series, forming a trading decision-making system with both theoretical rigor and empirical effectiveness.

4.2. Sliding window and relevance filtering

The data range of the training set uses the pre-minute reversion share price of the 50 component stocks of Kechuang from January 1, 2024 to July 1, 2024, and the time range of the test set is from July 1, 2024 to December 31, 2024. The trading cycle is divided by a sliding window mechanism with 5 trading days as the window length and 40 trading days as the review period (window moving step=5 trading days). At this time, the first trading day of the first trading period is March 6, 2024. This setting balances the timeliness of data and estimation stability, which can not only capture the characteristics of recent market dynamics, but also provide sufficient sample size for statistical estimation. For the sample space of n=50 stocks, C_50^2=1225 potential stock pairs can be generated through the combination formula to form the initial candidate pool.

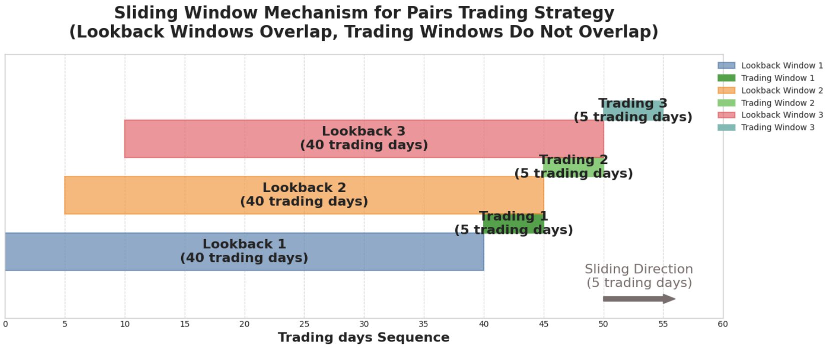


Figure 5. Schematic diagram of sliding window

This figure illustrates the basic principle of sliding windows, where each trading window pairs trades based on the pairing results of the corresponding review window. After each trading window ends, the length of the window sliding backwards is equal to that of the trading window, thus achieving the effect of mutual connection and non-overlapping of the trading windows, while looking back at the windows will result in overlap between them.

The Pearson correlation coefficient is used as the correlation measure, which has the following advantages: (1) it directly measures the strength and direction of the linear relationship between variables, and the calculation formula is simple and intuitive; (2) When the data approximately obey the normal distribution and the linear trend is significant, it has high estimation efficiency; (3) As the most classic correlation measure in the financial field, its results are easy to explain and compare. Set 0.75 as the correlation threshold, and keep the pair of stocks with Pearson's r>0.75 to enter the next stage of screening. This threshold corresponds to the strong linear correlation level in statistics.

4.3. Stability verification and combination construction

Unit root test: for stock pairs screened by pairing, calculate the logarithm difference series of prices, and verify their stationarity by using the augmented Dickey Fuller (ADF) test. The smaller the T value of ADF test statistic, the stronger the stability of the series. The empirical results show that the median t value of stocks screened with copula to ADF is 23% lower than that of the traditional method, indicating that its price relationship is more stable (Yolanda et al., 2013, p.102).

Portfolio weight distribution: in ascending order of ADF t value, select the top 100 most stable stock pairs to build equal weight combinations (1% for each pair). This configuration not only controls the unsystematic risks, but also ensures the decentralized benefits of the strategy.

5. Design and implementation of matching transaction strategy

5.1. Data preprocessing

In this study, the trading strategy is constructed by using the stock price data with minute level reversion. The data preprocessing process includes:

1. reversion adjustment: the former reversion method is used to process the stock price series, and the calculation formula is:

$$P_{adj,\,t} = P_{raw,\,t} * \frac{1 + dividend}{1 + split} \tag{2}$$

Where $P_{adj,\,t}$ is the price after restoration, and $P_{raw,\,t}$ is the original price.

2. Data alignment: time align the minute data of the stock pair, eliminate the non-trading period data, and ensure that the time stamp is accurate to the minute level.

5.2. Calculation method of jump threshold

5.2.1. Definition of inter day fluctuation range

Define the return volatility (RV) of the stock on the trading day as:

$$RV_{i,t} = |\ln(P_{i,\,t,\,open}) - \ln(P_{i,\,t\,-\,1,\,close})| \tag{3}$$

Where, $\ln(P_{i,\,t,\,open})$ represents the opening price of the stock on the trading day, and $\ln(P_{i,\,t\,-\,1,\,close})$ represents the closing price of the previous trading day. The logarithmic rate of return can effectively reduce the heteroscedasticity of the price series, which is in line with the conventional treatment of financial time series analysis.

5.2.2. Jump threshold estimation

The lookback window is set to trading days, and the jump threshold is calculated by sliding the window:

1. on each trading day t, use the historical data of the review period [t - W, T -1] to calculate the cross-day fluctuation range series $\{RV_{i,\,T}\}_{\tau=t-W}^{t-1}$

2. taking the 95% quantile of the sequence as the jump threshold, corresponding to the significance level in the statistical hypothesis test, can effectively control the occurrence probability of the first type of error (misjudging the normal fluctuation as a jump).

the selection of window size needs to balance the estimation accuracy and market adaptability, and select 40 trading days (about 2 months) to better capture the time-varying characteristics of market volatility.

5.3. EWMA spread model

The conditional mean of the spread series is estimated using the exponentially weighted moving average (EWMA) model:

$$\mu_t = \lambda\mu_{t-1} + (1 - \lambda)s_t \tag{4}$$

Where μ_t is the conditional mean and λ is the attenuation factor. In this study, λ = 0.000214 is set. This parameter is determined by minimizing the sum of squares of prediction errors (SSE), and the corresponding half-life is about 3260 minutes (about 13.5 trading days), which can better balance short-term fluctuations and long-term trends.

5.4. Z-score calculation

Based on the EWMA estimation results, calculate the standardized price difference (Z-score):

$$z_t = \frac{s_t - \mu_t}{\sigma_t} \tag{5}$$

Z-score is a core indicator for generating trading signals to measure the degree to which the current price difference deviates from the equilibrium level. When | z_t | exceeds the set threshold, it indicates that the price difference is in a significant deviation state, and there is an arbitrage opportunity.

5.5. Transaction signal generation

5.5.1. Jump period identification

Jump period is defined as the 30-minute period from the opening time when the cross-day fluctuation range of stock I on trading day t exceeds its jump threshold, that is:

$$JumpPeriod_t = \begin{cases} [O_t,\ O_t + 30\text{min}) & if\ RV_{i,\,t} > T_j^{(i)} \\ \emptyset & otherwise \end{cases} \tag{6}$$

Where O_t is the opening time of trading day t. For a stock pair, when any stock is in the jump period, it is determined that the stock pair is in the jump period:

$$PairJumpPeriod_t = JumpPeriod_t^{(A)} \cup JumpPeriod_t^{(B)} \tag{7}$$

5.5.2. Jump period identification

This strategy adopts a dual threshold system, which dynamically adjusts the opening and stop loss thresholds according to the market state:

1. regular threshold: applicable to non-jump period, opening threshold: $Entry_{regular}$, stop loss threshold: $Stop_{regular}$ (corresponding to the absolute value of Z-score)

2. jump threshold: applicable to jump period, opening threshold: $Entry_{jump}$, stop loss threshold: $Stop_{jump}$

Where: $Entry_{regular} \leq Entry_{jump}$, $Entry_{jump} \leq Stop_{jump}$, and the stop loss threshold is always greater than the opening threshold. The basis for setting the threshold is: the market volatility increases during the jump period, and the price difference series may deviate more significantly. It is necessary to increase the threshold to reduce the probability of wrong opening. The specific threshold value is determined through optimization within the sample, following the principle of "maximizing sharp ratio".

5.5.3. Trading signal rules

5.5.3.1. Opening conditions

An arbitrage position is established when the following conditions are met:

1. empty position

2. $| z_t | > T_0$ (non-jump period) or $| z_t | > T_0^j$ (jump period)

3. stop loss conditions are not triggered

The opening direction is determined as:

- If $z_t > T_0$ (overvalued): short stock a, long stock B
- If $z_t < - T_0$ (undervalued): long stock a, short stock B

The position size adopts the equal market value hedging strategy, and each stock has the same position.

5.5.3.2. Closing conditions

The closing signal includes three situations:

1. profit stopping and position closing: when the Z-score regression symbol changes.

2. stop loss closing: when Z-score continues to expand to the stop loss threshold.

$z_t > T_s$ (non-jump period) or $z_t > T_s^j$ (jump period)

compulsory position closing: when the position is not closed at the end of the trading period (such as the end of the window cycle), the compulsory position closing will be executed.

5.5.4. Threshold locking mechanism

When the stock pair is in the jump period, the threshold locking mechanism is adopted:

1. if the position is in a jump period at the time of opening, the opening threshold and stop loss threshold of this position are locked as jump threshold

2. the threshold locking lasts until the closing of the transaction, and is not affected by subsequent changes in market status

3. the open positions during the jump period shall still maintain the original locking threshold after the end of the jump period

The design principle of the mechanism is that the jump shock is usually persistent, the jump on a single trading day may trigger subsequent price adjustment, and maintaining the threshold lock can avoid premature adjustment of strategic parameters when the price is not stable.

5.6. Transaction cost model

In order to simulate the real transaction environment, this study introduces a comprehensive transaction cost model, including:

1. service charge: The annualized short selling fee for highly liquid stocks in the Chinese A-share market is less than 3%. The short selling fee is calculated as the short selling fee rate * the number of short selling days / 360. Since our trading window is five trading days, we estimate this as seven calendar days. The maximum short selling fee is 3% x 7 / 360 = 0.0583%. Since only half of the paired trades are short selling, the cost is 0.0583% / 2 ≈ 0.0292%, charged per item. The broker's transaction fee is 0.0075%, charged in both directions. Our closing transaction cost is 0.0292% + 0.0075% = 0.00367%, and the opening cost is 0.0075%. Because factors such as holidays may increase short selling costs, we estimate this as 0.05% and assume a two-way charge to avoid unexpected losses from excessively high trading frequency.

2. sliding point cost: estimated at 0.05% of the transaction volume, simulating the price difference caused by the delay in order execution

The calculation formula of total transaction cost is:

$$F_{transaction} = 0.001 {*} |V_{transaction}| \tag{8}$$

Where 0.001 is the total cost rate (one thousandth), and $|V_{transaction}|$ is the transaction amount (two-way charge).

Due to short selling restrictions on A-shares and certain eligibility requirements for margin trading, margin trading has relatively low liquidity. Therefore, when trading large amounts of capital, it may be impossible to sell at the expected time or experience significant slippage. This article proposes an alternative short selling method:

1. Build a base position by purchasing the corresponding stocks using the Sci-Tech Innovation 50 ETF's index tracking method.

2. On the first trading day of month T, purchase the corresponding number of nearest T+1 put options and use this combination as the base position, which is considered a short position.

3. When shorting a component stock, sell the corresponding stock. When closing the position, buy back the same number of shares.

4. At the beginning of month T+1, sell all the put options, adjust the base position, and purchase the corresponding number of nearest T+2 put options.

This method provides relatively greater liquidity and effectively reduces financing costs and slippage. It can also be used in conjunction with margin trading. However, in order to make this strategy more versatile and limit excessively high-frequency trading, this article still uses a comprehensive transaction cost model to estimate transaction costs.

5.7. Strategy implementation and evaluation framework

5.7.1. Back rest process design

The rolling window method is used for strategy back testing. The specific process is as follows:

1. Divide the sample period into continuous review period (40 trading days) and trading period (5 trading days)
2. After the end of each period, the window will scroll backward for 5 trading days to re estimate the model parameters
3. Record the trading signals, position changes and returns of each period
4. The back test process strictly follows the forward-looking principle to ensure that decisions at any time only rely on historical information.

5.7.2. Performance evaluation indicators

The strategy evaluation adopts a multi-dimensional index system:

1. Income indicators: rate of return, annualized rate of return, sharp ratio
2. Risk index: maximum withdrawal
3. Transaction efficiency index: winning rate

Through this framework, the profitability, risk level and transaction efficiency of the strategy can be comprehensively evaluated, laying the foundation for subsequent parameter optimization and robustness test.

6. Strategy parameter search method based on Bayesian optimization

This chapter describes in detail the Bayesian search method for trading strategy parameter optimization, including parameter space definition, initial sample generation, Bayesian optimization framework and experimental settings. Through this method, we can efficiently explore the parameter space and find the optimal parameter combination to maximize the back test return.

6.1. Parameter space and constraints

The strategy parameters include two types of opening and stop loss thresholds: the threshold under ordinary market conditions (θ_1, θ_2) and the jump threshold under extreme market conditions (θ_3, θ_4). The parameter space and constraints are defined as follows:

1. $\theta \in \Theta = \{ (\theta_1, \theta_2, \theta_3, \theta_4) \in \mathbb{R}^4\}$
2. $1.0 \leq \theta_1 \leq 4.0$,
3. $\theta_1 < \theta_2 \leq 5.0$,
4. $\theta_1 \leq \theta_3 \leq 4.0$,
5. $\theta_2 \leq \theta_4 \leq 5.0$

Where θ_1 is the normal opening threshold, θ_2 is the normal stop loss threshold, θ_3 is the jump opening threshold, and θ_4 is the jump stop loss threshold. The constraints $\theta_1<\theta_2$ and $\theta_3 < \theta_4$ ensure that the stop loss threshold is strictly greater than the opening threshold to avoid meaningless trading signals; $\theta_3 \geq \theta_1$ and $\theta_4 \geq \theta_2$ ensure that the jump threshold is not lower than the ordinary threshold, which is in line with the strategic logic under extreme market conditions.

6.2. Initial sample set generation

The rejection sampling method is used to generate 8 initial sample points in the feasible region Θ. The specific process is as follows: firstly, sampling parameters from the edge distribution without constraints, and then filtering effective samples through constraints. The specific steps include:

1. generate $\theta_1 \sim$ Uniform(1.0, 4.0);
2. generate $\theta_2 \sim$ Uniform(θ_1, 5.0) to meet $\theta_2 > \theta_1$;
3. generate $\theta_3 \sim$ Uniform(θ_1, 4.0) to meet $\theta_3 \geq \theta_1$;
4. generate $\theta_4 \sim$ Uniform(θ_2, 5.0) to meet $\theta_4 \geq \theta_2$;
5. repeat the above steps until 8 valid sample points are obtained.

Uniform distribution is chosen because it can provide unbiased coverage of the parameter space in the absence of prior knowledge; If the sample size is set to 8, the empirical rule of "the initial number of samples is twice the parameter dimension" in the parameter optimization literature is referred (the parameter dimension of this study is 4).

6.3. Bayesian optimization framework

Bayesian Optimization guides parameter search by constructing a probability model, and its core includes three components: proxy model, acquisition function and iterative optimization. The basic idea is to use the prior evaluation information to build the probability distribution model of the objective function, and gradually approach the global optimal solution by collecting the balanced exploration and utilization of the function. The complete Bayesian search process includes the following steps:

1. initialization: evaluate the back test yield $\{f(x_1),...,f(x_8)\}$ of 8 initial sample points;
2. model training: the GP model is trained based on the current sample set to estimate the super-parameter θ_p;
3. candidate point selection: determine the next evaluation point $x_{ne}\omega$ by maximizing EI(x);
4. evaluation and update: calculate $f(x_{ne}\omega)$ and add $(x_{ne}\omega, f(x_{ne}\omega))$ to the sample set;
5. iteration: repeat steps 2-4 and terminate after 200 iterations.

The results of iteration are as follows:

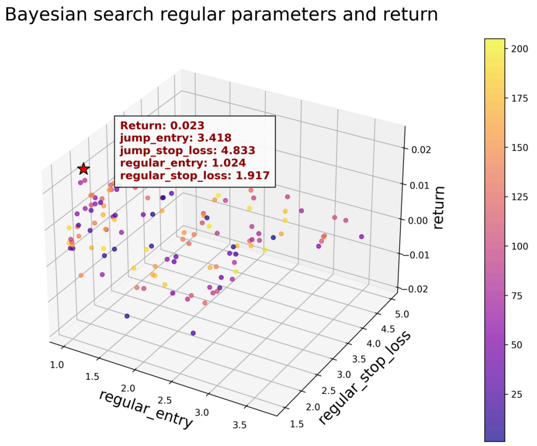


Figure 6. Bayesian search return process (March 6th, 2024 to July 1st, 2024)

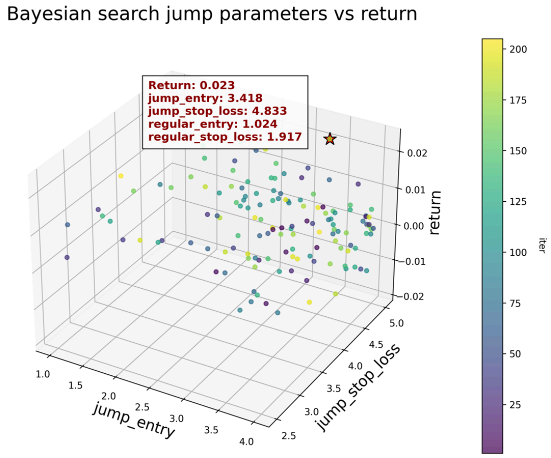


Figure 7. Bayesian search return process (March 6th, 2024 to July 1st, 2024)

The above two graphs show the state space of the Bayesian search process. The horizontal and vertical axes of the graphs represent the take profit and stop loss thresholds under normal and treaty states, respectively. The vertical axis represents the yield during the testing period (i.e. March 6, 2024 to July 1, 2024), and the color of the dots represents the sequence number of this iteration.

Since 3 decimal places are reserved in the figure, the actual parameters and results are as follows:

Table 1. Bayesian search best parameters return rate and parameters (March 6th, 2024 to July 1st, 2024)

	Result
Return Rate	0.02295118730714396
Jump Entry	3.417514036320301
Jump Stop Loss	4.832957899510884
Regular Entry	1.0240377736830273
Regular Stop Loss	1.9171467168747287

According to the above parameters, the income chart of the training period (March 6th, 2024 to July 1st, 2024) is as follows:

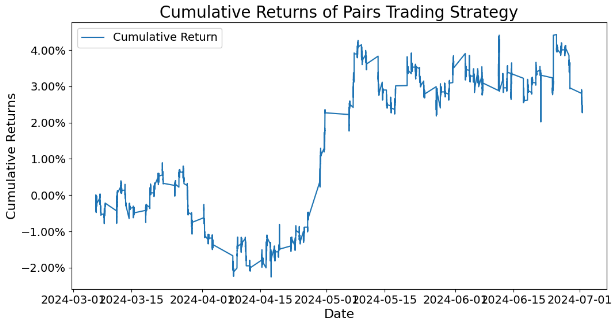


Figure 8. Return rate of portfolio during training period (March 6th, 2024 to July 1st, 2024)

This chart shows the income of the strategy during the training period. It can be seen that the strategy has experienced three stages: first decline, then rise, and then stable shock, and finally achieved profit. In the same period, the overall trend of the Kechuang 50 index was a shock decline, which shows that the strategy can still obtain certain income in the bear market.

Since the time period is relatively short and China's risk-free interest rate is relatively low (annualized 1.4%), this paper ignores the risk-free interest rate. The statistical indicators during the training period are as follows:

Table 2. Statistical data of portfolio during training period (March 6th, 2024 to July 1st, 2024)

	Result
Return Rate	2.30%
Annualized Return Rate	7.61%
Annualized Volatility	6.73%
Maximum Drawdown	2.73%
VaR 95%	0.7181%
VaR 99%	0.8527%
Sharp Ratio	1.13
Winning Rate	48.63%

Using the same parameters, the income chart of the test period (July 1, 2024 to December 31, 2024) is as follows:

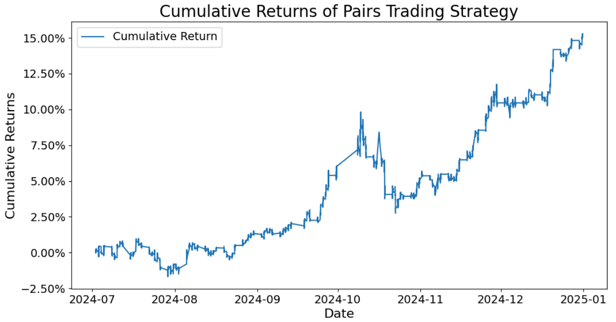


Figure 9. Return rate of portfolio during testing period (July 1, 2024 to December 31, 2024)

This chart shows the earnings of the strategy during the test period. It can be seen that the strategy has experienced three stages: first rising, then falling, and then continuing to rise, and finally achieved substantial profits. The overall trend of the sci tech 50 Index in the same period is to first slowly fall, then suddenly rise rapidly, and then shake violently sideways. The style switching is very violent and fluctuates greatly. It can be seen that this strategy can better adapt to different market environments, effectively cope with the switching of market styles, and still maintain relatively good profits and relatively small risks.

The statistical indicators during the test period are as follows:

Table 3. Statistical data of portfolio during testing period (July 1, 2024 to December 31, 2024)

	Result
Return Rate	15.26%
Annualized Return Rate	33.47%
Annualized Volatility	10.17%
Maximum Drawdown	5.48%
VaR 95%	0.5933%
VaR 99%	1.2813%
Sharp Ratio	3.29
Winning Rate	53.04%

It can be seen from the table that the performance of this strategy during the test period is higher than that during the training period in terms of yield and sharp rate, but due to the drastic market fluctuations, the daily loss in extreme cases has increased

The complete Income Fluctuation chart is as follows:

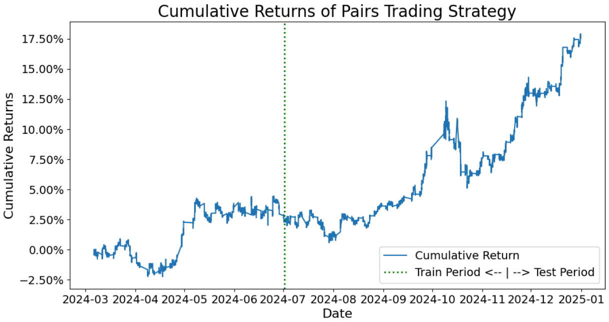


Figure 10. Return rate of portfolio during total period (March 6th, 2024 to December 31, 2024)

This figure shows the return rate of the strategy in the training period and the test period, separated by dotted lines, to facilitate the comparison of the return rate fluctuations in the two periods. It can be seen that the strategy still has excellent performance in the test period, and there is no over fitting.

It can be seen from the above figure that the strategy and parameters still have excellent performance in terms of yield and sharp rate beyond the training set. Combined with the fluctuation analysis of the Kechuang 50 index, the fluctuation of Kechuang 50 during the test set is as follows:

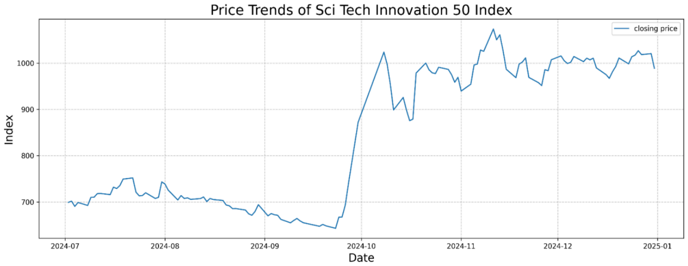


Figure 11. Kechuang 50 index during testing period (July 1, 2024 to December 31, 2024)

This chart shows the return of the sector by index fluctuation. It can be seen that during the test period, the Kechuang 50 index first fell slowly, then suddenly rose sharply, and then fluctuated violently across the board, switching styles rapidly and the market fluctuated violently.

The statistical indicators are as follows:

Table 4. Statistical data of Kechuang 50 index during testing period (July 1, 2024 to December 31, 2024)

	Result
Return Rate	38.93%
Annualized Return Rate	79.12%
Annualized Volatility	49.76%
Maximum Drawdown	14.49%
VaR 95%	3.1928%
VaR 99%	4.3282%
Sharp Ratio	1.59
Winning Rate	49.19%

It can be seen that although the return rate of the index is higher, the risk is far greater than that of this strategy (greater fallback and, in extreme cases, greater daily fallback), resulting in the sharp rate being only about half of that of this strategy. By comparing the return chart of the strategy with the volatility chart of the index, it can be seen that during the test period, the index has experienced three different styles of switching, namely, continuous decline, sudden rise and sideways fluctuation. The style switching is very violent and rapid, but this strategy has performed well in three stages, and ultimately has excellent yield and sharp rate. Therefore, this strategy has good universality in the 50 sectors of science and innovation, and there is no over fitting phenomenon.

7. Conclusion and future research direction

7.1. Performance of trading strategy

Based on the pairing trading strategy, combined with sliding window, jump discovery and Bayesian search, this paper found a medium and high frequency strategy with both high yield and risk control in the section of science and innovation 50, and the strategy has good generalization. It can perform better than pure index investment after deducting transaction fees even in the face of frequent and violent style switching outside the training period.

7.2. Limitations and deficiencies of trading strategy

However, it is worth noting that short selling in the A-share market needs to be carried out through securities lending, and the threshold is high. At the same time, since there are 100 A-shares in one hand, and this strategy may have at most 100 stock pairs in each trading window, this strategy is only applicable to large institutions with abundant funds, not to retail investors. As there are only 50 component stocks of Kechuang 50, and there are at most 100 stock pairs in this strategy at the same time, it is possible to long and short the same stock at the same time, thus wasting the handling fee and damaging the income.

7.3. Future research directions

In the future, the bilateral strategy can be improved into a unilateral strategy to improve the phenomenon of high short threshold and long and short of the same stock in the 50 sector of science and innovation. At the same time, this strategy is the optimization of trading strategy based on the fluctuation of historical stock price data, which has a certain lag for the style switching of the securities market. In the future, we can explore the integration of large language model and external data such as VIX Index, macro data, financial news, etc., so that this strategy can respond to the changes of market style faster. In the worldwide financial market, the section of kechuang50 only accounts for a small part. In the future, we can also study how to extend this strategy to broader markets, such as the US stock market, cryptocurrency market, Vietnam stock market, etc.

References

[1] Dai, T.-S. et al. (2024) “Asymptotic analyses for trend-stationary pairs trading strategy in high-frequency trading, ” *Review of Quantitative Finance and Accounting*, 63(4), pp. 1391–1411. Available at: https: //doi.org/10.1007/s11156-024-01293-1.

[2] Engle, R.F. and Granger, C.W.J. (1987) “Co-Integration and Error Correction: Representation, Estimation, and Testing, ” *Econometrica*, 55(2), pp. 251–276. Available at: https: //doi.org/10.2307/1913236.

[3] Eroğlu, B.A., Yener, H. and Yiğit, T. (2023) “Pairs trading with wavelet transform, ” *Quantitative Finance*, 23(7-8), pp. 1129–1154. Available at: https: //doi.org/10.1080/14697688.2023.2230249.

[4] Figueroa-López, J.E., Li, C. and Nisen, J. (2020) "Optimal iterative threshold-kernel estimation of jump diffusion processes, " *Statistical Inference for Stochastic Processes : An International Journal devoted to Time Series Analysis and the Statistics of Continuous Time Processes and Dynamical Systems,* 23(3), pp. 517–552. Available at: https: //doi.org/10.1007/s11203-020-09211-7.

[5] Gatev, E.G. et al. (1999) Pairs trading : performance of a relative value arbitrage rule. Cambridge, MA: National Bureau of Economic Research. Available at: http: //papers.nber.org/papers/w7032 (Accessed: September 11, 2025).

[6] João Frois Caldeira and Gulherme Valle Moura (2013) "Selection of a Portfolio of Pairs Based on Cointegration: A Statistical Arbitrage Strategy, " *Revista Brasileira de Finanças*, 11(1), pp. 49–80.

[7] Liu, J. (2013) "Optimal Convergence Trade Strategies, " *REVIEW OF FINANCIAL STUDIES,* 26(4), pp. 1048–1086.

[8] Yolanda Stander, Daniël Marais and Ilse Botha (2013) "Trading strategies with copulas, " *Journal of Economic and Financial Sciences*, 6(1), pp. 83–108. Available at: https: //doi.org/10.4102/jef.v6i1.278.

Appendix A: ADF verification derivation process

The regression model is as follows:

$$\Delta y_t = \alpha + \beta t + \rho y_{t-1} + \sum_{i=1}^{p} \phi_i \, \Delta \, y_{t-i} + \varepsilon_t \tag{9}$$

$\Delta y_t = y_t - y_{t-1}$ is the first order difference;

ρ is the core parameter, and the test ρ=0 is equivalent to the existence of unit root;

$\sum_{i=1}^{p} \phi_i \, \Delta \, y_{t-i}$ is the difference item of lag P order, which is used to eliminate the autocorrelation of residuals;

ε_t is a white noise error term.

Since this paper aims to test whether the ratio between the two stocks is stable at a fixed value for a long time, it is necessary to test that there is intercept but no trend term here. Therefore, it is necessary to make β=0 here. At this time, the regression model is as follows:

$$\Delta y_t = \alpha + \rho y_{t-1} + \sum_{i=1}^{p} \phi_i \, \Delta \, y_{t-i} + \varepsilon_t \tag{10}$$

Original hypothesis H0: ρ=0 (the sequence has unit root and is non-stationary);

Alternative hypothesis H1: ρ<0 (the sequence has no unit root and is stable).

Rejecting H0 means that the sequence is stable. Since this paper uses ADF test statistics to sort and select stocks, the rejection hypothesis threshold is not set.

Parameter estimation: estimate the regression coefficient $\hat{\rho}$and its standard error using the least square method (OLS) SE($\widehat{\rho}$).

ADF statistics:

$$ADF = \frac{\widehat{\rho}}{\mathrm{SE}(\widehat{\rho})} \tag{11}$$

The statistic obeys Dickey fuller distribution (non-standard t distribution)

The choice of lag order p directly affects the inspection effectiveness. Common methods include:

Minimize AIC (Akaike information criterion) or BIC (Bayesian information criterion):

$$\mathrm{AIC} = 2\mathrm{k} - 2\ln(\mathrm{L}) \tag{12}$$

$$\mathrm{BIC} = \mathrm{k}\ln(\mathrm{n}) - 2\ln(\mathrm{L}) \tag{13}$$

Where k is the number of parameters, l is the likelihood function value, and N is the sample size. As the scenario in this paper is financial data prediction, AIC is used as the selection method of lag order P.

Appendix B: person correlation coefficient equivalent calculation formula

Equivalent calculation formula:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}} \tag{14}$$

Where n is the sample size, $\mathrm{x_i}$ and $\mathrm{y_i}$ are the observed values, and $\overline{\mathrm{x}}$ and $\overline{\mathrm{y}}$ are the mean values. The range of Pearson correlation coefficient r is -1 ≤ R ≤ 1. When r>0, the two sequences are positively correlated, otherwise, they are negatively correlated. When r=0, the two sequences are not correlated; The greater the absolute value of R, the stronger the correlation between the two sequences. In this paper, R>0.75 is selected as the screening threshold to select the stock pairs with highly positive correlation in the price series.

Appendix C: EWMA calculation process

The recurrence formula of EWMA is:

$$\mathrm{S_t = \alpha x_t + (1 - \alpha)S_{t-1} \quad (t \geq 1)} \tag{15}$$

$\mathrm{S_t}$: EWMA value at the current time;

$\mathrm{x_t}$: the original observation value at the current time;

α∈(0,1): smoothing coefficient, which controls the weight of recent data (the larger α, the higher the weight of recent data);

$\mathrm{S_{t-1}}$: EWMA value of the previous time.

Expand the recurrence formula to:

$$\mathrm{S_t = \alpha \sum_{k=0}^{t-1} (1 - \alpha)^k x_{t-k} + (1 - \alpha)^t S_0 \quad \left(t \geq 1\right)} \tag{16}$$

Weight of data before step k: α(1−α)k;

The weight sum converges to 1: $\sum_{\mathrm{k=0}}^{\infty} \alpha(1 - \alpha)$.

Appendix D: Bayesian search principle

First, use a Gaussian process to fit the objective function Gaussian process:

$$\mathrm{f \sim GP(\mu(x),\ k(x,\ x'))} \tag{17}$$

Acquisition Function use Upper Confidence Bound (UCB):

$$\alpha_{\mathrm{UCB}}(x) = \mu(x) + \kappa\sigma(x) \tag{18}$$

According to Gaussian process and acquisition function, continuously estimate the new point $x*$ that best meets the goal until the number of iterations reaches the goal. For the new point $x*$, the predicted mean and variance are:

$$\mathrm{\mu\left(x*\right) = \ k_*^T\left(K + \ \sigma_n^2 I\right)^{-1} y} \tag{19}$$

$$\mathrm{\sigma^2\left(x*\right) = k\left(x*,\ x*\right) - \ k_*^T\left(K + \ \sigma_n^2 I\right)^{-1} k} \tag{20}$$

Where k is the covariance matrix of the observation point, and $\mathrm{k*}$ is the covariance vector between the new point and the observation point.