▪ Journal of Competitiveness

Evaluating a Stock Market Pairs Trading Strategy Using a Dependent Functional Data Cointegration Method

Danni Wanga and Zhifang Sua

Abstract:
Pairs trading, recognized as an effective high-frequency investment strategy to mitigate risk, has been widely employed in statistical arbitrage. Traditional pairs trading approaches often impose strict constraints on asset selection, and in view of the functional and dependent characteristics of financial market data, this study proposes an innovative pairs trading strategy based on a dependent functional cointegration model. Specifically, we first utilize a dependent functional cointegration test to identify contract pairs exhibiting cointegration relationships, then establish a signal mechanism and threshold setting for the trading phase, and finally carry out arbitrage tests in stock markets. The results show that, compared with traditional methods, this approach not only respects the stochastic nature of data but also operates under a nonparametric framework, enabling accurate recognition of dependence and dynamic changes in high-frequency data. It thereby precisely captures short-term mean reversion trends, achieves higher trading frequency, and yields significantly improved returns. Consequently, this method offers trading institutions a practical tool for constructing a universal quantitative trading system in fiercely competitive markets, contributing to the overall competitiveness and efficiency of the financial sector.

Keywords: *Dependent functional data, Co-integration test, Pair trading*

JEL: C812，F830

Article history: Received: November 2024; Accepted: August 2025; Published: December 2025

1. INTRODUCTION

Amid the accelerating globalization of financial markets and rapid advancements in information technology, high-frequency trading has shown a remarkable capacity to capture microstructural price fluctuations, executing transactions within extremely short time frames. This capability not only expands profit opportunities for market participants but also serves as a crucial means of enhancing overall market competitiveness. Pairs trading, a commonly adopted and relatively robust statistical arbitrage strategy, identifies two assets or indices with a stable long-term price relationship; when the price spread between them deviates from a predefined threshold, the strategy shorts the overvalued asset and goes long on the undervalued asset, and then closes both positions once the spread reverts to its historical mean to realize arbitrage returns. Especially in the context of rapidly evolving high-frequency trading, accurately selecting asset

https://doi.org/10.7441/joc.2025.04.08

pairs in a high-dimensional and dynamically changing market environment, as well as determining the optimal trading threshold, has become the linchpin of successful statistical arbitrage (MacKinlay & Ramaswamy, 1988).

Traditional pairs trading research often employs correlation coefficient approaches, the minimum distance method (Gatev et al., 2006), and the random spread method (Elliott et al., 2005) to select asset pairs. However, when dealing with nonstationary and highly volatile high-frequency data, these methods frequently lead to spurious regressions or overlook critical dynamic information. Engle and Granger (1987) introduced the E-G two-step method as a feasible cointegration test capable of effectively detecting long-term equilibrium relationships among assets. Vidyamurthy (2004) extended this approach to stock market pairs trading by selecting equities with stable long-run relationships and employing mean reversion to capitalize on abnormal spread deviations, thereby mitigating the limitations of the aforementioned traditional methods while reducing transaction costs and maximizing profits. Literature reviews by Huck and Afawubo (2015), Rad et al. (2016), and Mario et al. (2018) similarly reveal that cointegration-based pairs trading strategies more accurately retain the original data characteristics. Building upon these findings, some scholars (e.g., Vychytilová & Jana, 2015) have applied multivariate statistical methods to investigate cross-border financial market linkages, further evidencing inter-market connections. Meanwhile, Huang and Martin (2019), RamoRequena et al. (2017), and others have incorporated cointegration models into cross-market statistical arbitrage, discovering that, irrespective of market volatility, stable excess returns can be obtained when assets exhibit cointegration relationships.

As data frequency continues to rise, relying on traditional low-frequency modelling methods makes it difficult to accurately capture price dynamics and cross-market linkages (Madden, 2012), thereby undermining the sustainability of trading strategies. Due to the high volatility and instability of financial markets, historical data cannot fully reflect future market movements. Leung et al. (2001) emphasize that, under such conditions, incorporating forecasting mechanisms into trading models and adjusting them in real time is essential to achieve better performance in highly volatile markets. High-frequency financial data often exhibit high dimensionality, strong noise, and pronounced temporal dependence, making it challenging for conventional econometric models to effectively capture continuously evolving high-frequency trading processes.

By contrast, functional data analysis (FDA) projects discrete observations into a functional space and employs dimensionality-reduction techniques such as principal component decomposition, thereby more flexibly revealing the temporal and spatial dynamics of financial variables. Building on the work of Ramsay (1991, 2002), Huang (2001), and Müller (2011), a

https://doi.org/10.7441/joc.2025.04.08

comprehensive theoretical framework has been established. On this basis, Dauxois et al. (1982) utilized functional principle component analysis (FPCA) to convert infinite-dimensional characteristic vectors into finite-dimensional score vectors. Boubaker et al. (2021) applied FDA to develop novel trading rules, reconstructing time-series momentum portfolios. In recent years, FDA has also demonstrated greater applicability in empirical analysis and has been extensively employed in fields such as meteorology, economics, and biomedicine (Tsay, 2016; Ceroveck et al., 2019; Das et al., 2019).

However, existing research based on functional data analysis typically assumes that functional data follow an independent and identically distributed process, neglecting the dependence and interlinkages often seen in real-world asset prices. To address this limitation, Hörmann et al. (2010), Horváth et al. (2012), and Kokoszka et al. (2017) introduced dependent functional data analysis, which incorporates a long-run covariance function to capture the dependency structure among data series. This approach yields more accurate eigenvalues and eigenfunctions, thereby delivering more robust estimation results in the noise-heavy context of high-frequency financial data. Nevertheless, the estimation of the long-run covariance function inevitably involves selecting both a kernel function and a bandwidth—choices that, if made incorrectly, can introduce notable estimation errors.

In light of these limitations, this study proposes a new high-frequency pairs trading strategy based on a dependent functional cointegration model, conducting empirical tests of statistical arbitrage in stock markets. By converting intraday high-frequency prices into dependent functional data and reconstructing functional curves from discrete trading prices, this approach identifies cointegrated asset pairs that are integrated of the same order. Next, the strategy defines the trading cycle, entry and exit signals, and relevant parameters, while incorporating a rolling regression mechanism for dynamic adjustments. Consequently, it offers a more precise depiction of short-term mean reversion and long-term stability among assets. Empirical results indicate that this strategy not only enables higher intraday trading frequency and superior returns, but also effectively reduces the noise from high-dimensional data, thereby facilitating timely responses and optimal decision-making in the highly competitive landscape of international financial markets.

The remainder of this paper is organised as follows. Section 2 introduces the model construction. Section 3 set up a matching trading process. Section 4 presents and discusses the empirical results. Finally, Section 5 summarises the main conclusions.

2. METHODS

https://doi.org/10.7441/joc.2025.04.08

The primary task in constructing a pairs trading strategy is to screen asset portfolios. To filter the noise in raw data, this study reconstructs and smooths discrete observations into functional curves from a functional perspective. In view of the dependence typically observed in financial market data, we propose a novel long-run covariance estimator based on the untruncated Bartlett kernel. This corrects the covariance function assumed under the condition of independence and identical distribution, thereby reconstructing discrete data without having to artificially select a kernel function and optimal bandwidth. We then extend the traditional cointegration model to a dependent functional cointegration model, identifying asset pairs that exhibit stable long-term relationships for the subsequent pairs trading stage.

2.1. Reconstruction of Dependent Functional Data

A random variable originating $X_i(t)(t=1,2,...,T)$ from a square-integrable Hilbert space $H=L^2[0,T]$, with each sample curve containing T discrete observations, is referred to as a functional dataset. By appropriately representing the latent structure of these discrete observations as a smooth curve or continuous functional process $X_{ij}=X_i(t_j)$, $t\in T, i=1,2,\cdots,N$, one obtains what is called the functionalized process.

When $X_n(t)$ fails to satisfy the I.I.D. assumption and exhibits dependence on its lagged values, reference can be made to Hörmann and Kokoszka (2010) to adjust the covariance function initially established under the assumption of independence and identical distribution. Specifically, the long-run covariance function is used in place of the I.I.D.-based covariance to refine the functional data. The sample mean and variance are expressed as follows:

$$Var\left[\bar{X}_N\right]=Var(N^{-1}\sum_{n=1}^{N}X_n)=N^{-2}\sum_{n=1}^{N}Var\left[X_N\right]=N^{-2}\sum_{m,n=1}^{N}Cov(X_m,X_n) \quad (1)$$

Let the time interval be $h=m-n$. Then the equation can be rearranged as follows:

$$\mathrm{Var}\left[\bar{X}_N\right]=N^{-2}\sum_{h=-N+1}^{N-1}(N-|h|)\gamma_h=N^{-1}\sum_{h=-N+1}^{N-1}\left(1-\frac{|h|}{N}\right)\gamma_h \quad (2)$$

By the dominated convergence theorem, if $\sum_{h=-\infty}^{\infty}|\gamma_h|<\infty$, we have $\lim_{N\to\infty}N\ \mathrm{Var}\left[\bar{X}_N\right]=\sum_{h=-\infty}^{\infty}\gamma_h<\infty$.

for any h, when $N\to\infty, h/N\to 0$ holds. Hence the time series $\{X_n\}$ satisfies:

https://doi.org/10.7441/joc.2025.04.08

$$\sqrt{N}\left(\bar{X}_N-\mu\right)\xrightarrow{d}N\left(0,\sum_{h=-\infty}^{\infty}\gamma_h\right) \tag{3}$$

The long-run covariance function is defined as:

$$\gamma=\gamma_0+\sum_{h=1}^{\infty}\left(\gamma_h+\gamma_{-h}\right) \tag{4}$$

The sample long-term covariance estimator for the sum of two independent variables s and t is defined as:

$$\gamma\left(s,t\right)=\gamma_0\left(s,t\right)+\sum_{h=1}^{\infty}\left[\gamma_h\left(s,t\right)+\gamma_{-h}\left(s,t\right)\right] \tag{5}$$

where $\gamma_h\left(s,t\right)$ and $\gamma_{-h}\left(s,t\right)$ are self-covariance functions of order H, which are defined as:

$$\begin{aligned}\gamma_h\left(s,t\right)&=Cov[X_i(s),X_{i+h}(t)]=E\left\{[X_i(s)-\mu(s)][X_{i+h}(t)-\mu(t)]\right\}\\\gamma_{-h}\left(s,t\right)&=Cov[X_i(s),X_{i-h}(t)]=E\left\{[X_i(s)-\mu(s)][X_{i-h}(t)-\mu(t)]\right\}\end{aligned}. \tag{6}$$

moreover, $h=0$, $\gamma_h\left(s,t\right)$ is the short-term covariance function under the I.I.D. And, the convergence relation holds:

$$\lim_{N\to\infty}N\,Cov\left[\bar{X}_N(s),\bar{X}_N(t)\right]=\lim_{N\to\infty}\sum_{h=-N+1}^{N-1}\left(1-\frac{|h|}{N}\right)\gamma_h\left(s,t\right)\to\sum_{h=-\infty}^{\infty}\gamma_h(s,t)<\infty. \tag{7}$$

When discrete observation data are collected, the kernel function method can be used to estimate the long-term covariance function as:

$$\begin{aligned}\hat{\gamma}(s,t)&=\hat{\gamma}_0(s,t)+\sum_{h=1}^{N-1}K(\frac{h}{q})[\hat{\gamma}_h\left(s,t\right)-\hat{\gamma}_{-h}\left(s,t\right)]\\\hat{\gamma}_h\left(s,t\right)&=\frac{1}{N-h}\sum_{i=1}^{N-h}[X_i(s)-\bar{X}_N(s)][X_j(t)-\bar{X}_N(t)]\end{aligned} \tag{8}$$

The optional kernel function is defines as $K(x)=1-|x|,|x|\le1$. Some scholars also choose the Newey-West kernel function to estimate the long-term covariance function (e.g.Kokoszka and Young, 2017), which is defined as:

$$\omega_m(h)=\begin{cases}1-\dfrac{h}{m},h\le m\\0,h>m\end{cases},m=4*(\frac{N}{100})^{\frac{2}{9}}$$

https://doi.org/10.7441/joc.2025.04.08

However, the foregoing methods for estimating the long-run covariance function all face the challenge of choosing a kernel function and an optimal bandwidth. An inappropriate selection can introduce large errors. To address this shortcoming, following Kiefer and Timothy (2002), this study proposes an untruncated Bartlett kernel-based long-run covariance estimator, effectively overcoming the aforementioned limitations. The Bartlett kernel long-run covariance estimator is defined as:

$$\hat{\gamma}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}(1-\frac{|i-j|}{N})v_i v_j' \quad . \tag{9}$$

Select the samples between the time point variables s and t from the above formula and convert the above formula into:

$$\gamma(s,t)=\gamma_0(s,t)+\sum_{h=1}^{\infty}[\gamma_h(s,t)+\gamma_{-h}(s,t)] \, . \tag{10}$$

The long-term covariance estimation formula of the sample between two variables is as follows:

$$\hat{\gamma}(s,t)=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}(1-\frac{|i-j|}{N})[X_i(s)-\bar{X}_N(s)][X_j(t)-\bar{X}_N(t)] \tag{11}$$

After estimating the long-term covariance of the sample, the principal component $\hat{\phi}_k$ and eigenvalue of the corresponding function $\hat{\lambda}_k$ can be calculated, and the principal component score of the function can be obtained as:

$$\hat{\xi}_{ik}=\int[X_i(t)-\hat{\mu}(t)]\hat{\phi}_k(t)dt \, . \tag{12}$$

Using the K-L expansion, the function can be further expressed as:

$$X_i(t)=\mu(t)+\sum_{k=1}^{\infty}\xi_{ik}\phi_k(t)\approx\hat{\mu}(t)+\sum_{k=1}^{K}\hat{\xi}_{ik}\hat{\phi}_k(t) \tag{13}$$

The Bartlett kernel-based long-run covariance estimator proposed in this study does not require manual selection of either a kernel function or a bandwidth, making it simpler and more logically consistent than previously documented covariance estimation methods. Moreover, it achieves greater accuracy compared to conventional functional data estimation approaches.

First, stationarity tests must be conducted on both the reconstructed dependent functional time series and their first-order differenced series to identify asset pairs integrated of the same order for subsequent pairs trading. In this study, we apply the stationarity testing method proposed

https://doi.org/10.7441/joc.2025.04.08

by Horváth et al. (2014) to select asset pairs integrated of the same order, which are then employed in the subsequent dependent functional cointegration tests.

2.2. Dependent function cointegration test

In traditional literature, the methods most frequently used for selecting paired asset combinations include the distance method, the random spread method, and the cointegration method. The distance method is a nonparametric pairs trading approach based purely on statistical criteria for selection. In contrast, the random spread method employs continuous time series models, applicable only to products or companies of the same type or brand listed in financial markets, and assumes that paired assets exhibit identical long-term mean returns. The cointegration method, as a parametric approach, considers features such as nonstationarity and dynamic relationships inherent in financial data, enabling an effective estimation of deviations between two asset combinations, and thus facilitating appropriate timing for statistical arbitrage. Vidyamurthy (2004) introduced cointegration theory into pairs trading by identifying pairs of stocks with stable equilibrium relationships. Arbitrage is then conducted based on mean reversion when the prices of paired assets deviate significantly from their long-term equilibrium, effectively avoiding the spurious regression problem that arises when regression models are directly applied to nonstationary data.

Assume that the asset pair portfolio X and Y are the same order integral dependent functional stationary variables, and obtain the principal component basis $\phi^X(t)$ and $\phi^Y(t)$ by using the long-term covariance function, and select the K_X and K_Y principal component functions according to the cumulative contribution rate. Samples of dependent function type can be approximately expressed as:

$$\begin{aligned} X_i^c(t) &= \mu_X(t) + \sum_{l=1}^{\infty} \xi_{il}^X \phi_l^X(t) \approx \mu_X(t) + \sum_{l=1}^{K_X} \xi_{il}^X \phi_l^X(t) \\ Y_i^c(t) &= \mu_Y(t) + \sum_{k=1}^{\infty} \xi_{ik}^Y \phi_k^Y(t) \approx \mu_Y(t) + \sum_{k=1}^{K_Y} \xi_{ik}^Y \phi_k^Y(t) \end{aligned} \tag{16}$$

The linear regression model is expressed as:

$$Y_i^c(t) = \beta_0(\mathrm{t}) + \int \beta(s,t) X_i^c(s) ds + \varepsilon_i(t), \quad i = 1, 2, \ldots, N \tag{17}$$

Where, $\beta_0(\mathrm{t})$ is the intercept function, $\varepsilon_i(t)$ is the error term with zero mean, and the linear

https://doi.org/10.7441/joc.2025.04.08

regression coefficient of financial function is expressed as:

$$\beta(s,t)=\sum_{l=1}^{K_x}\sum_{k=1}^{K_y}b_{lk}\phi_l^X(s)\phi_k^Y(t) \tag{18}$$

The cointegration test is extended to the framework of dependent functional data analysis based on the long-run covariance function. Following the logic of the E-G two-step procedure, a dependent functional cointegration test model is constructed to examine whether two integrated-of-order-one functional variables exhibit a cointegration relationship. First, the residual sequence is obtained through dependent functional cointegration regression, after which the stationarity of this residual sequence is tested to further determine the existence of cointegration between variables. The specific testing procedure is as follows:

The first step is to establish the dependent function type linear regression mode $Y_i^c(t)=\beta_0(t)+\int\beta(t,s)X_i^c(s)ds+\varepsilon_i(t)\ i=1,2,...,N$, and to obtain the cointegration regression and calculate the disequilibrium error function $e_i(t)$.

$$\begin{aligned}\hat{Y}_i^c(t)&=\hat{\beta}_0(\mathrm{t})+\int\hat{\beta}(t,s)\hat{X}_i^c(s)ds+\varepsilon_i(t)\\ e_i(t)&=Y_i^c(t)-\hat{Y}_i^c(t)\end{aligned} \tag{19}$$

The second step is to verify the disequilibrium error function $e_i(t)$ by using the dependent-function stationarity test. If $e_i(t)$ is stationary sequence, then variables are considered $X_i(t),Y_i(t)$ are first-order co-integration. In contrast, think of dependent functional variables $X_i(t),Y_i(t)$ There is no cointegration relationship between.

3. MATCHING TRANSACTION PROCESS

Drawing on the core logic of pairs trading and mean reversion, as well as hedged trading methods, this study references Gatev et al. (2006), Vidyamurthy (2004), and Caldeira and Moura (2013) to develop a dynamic hedging strategy. By exploiting mean reversion properties and focusing on stationary spread series to determine entry and exit timing, this approach rectifies the model gaps in the minimum distance and random spread methods. The specific trading workflow is as follows: first, dependent functional data analysis is applied to reconstruct smooth functional curves from discrete trading prices, thereby identifying cointegrated asset pairs that share the same order of integration; second, the trading cycle, entry and exit signal

https://doi.org/10.7441/joc.2025.04.08

mechanisms, and trading parameters are established; and finally, aiming for optimal returns, the cointegration-based pairs trading strategy is evaluated.

3.1. Clustering Analysis Method

Clustering analysis is a type of unsupervised learning within machine learning, which categorizes data purely based on intrinsic data characteristics without relying on any prior knowledge. It ensures objects within clusters exhibit high similarity, while objects across clusters display significant dissimilarity. Considering that one of the primary objectives of this study is to identify pairs of stocks exhibiting cointegration relationships for trading purposes, and given that trading decisions are closely related to the price spreads between stock pairs, we select the K-means clustering method based on numerical distances between functional curves. Specifically, by applying multivariate clustering to the functional principal component scores, the infinite-dimensional functional data clustering problem is effectively transformed into a finite-dimensional clustering problem. The primary objective of the K-means algorithm is to partition the dataset into K clusters, such that samples within the same cluster are highly similar, whereas samples between different clusters are notably dissimilar.

The specific steps are as follows: randomly select K initial center points from the observed sample curves. Then compute the numerical distances between each of the remaining sample curves and these K center curves. Based on the principle of minimum distance, assign each curve to the closest of the K clusters.

$$d(x,y)=\sqrt{\sum_{j=1}^{n}(x_j-y_j)^2}$$

Finally, calculate the mean curve of each cluster and treat these means as new cluster centers for the next iteration. Repeat the aforementioned process iteratively until there is no further improvement in the objective function.

$$E=\sum_{i=1}^{K}\sum(x-\bar{x})^2$$

The optimal number of clusters K is determined by referencing the Calinski and Harabasz (1974) CH index, while simultaneously considering the ratio of between-cluster sum of squares to the total sum of squares. Specifically, the optimal number of clusters q corresponds to the maximum value of the CH index.

$$CH(q)=\frac{trace(B_q/(q-1))}{trace(W_q/(n-q)}$$

https://doi.org/10.7441/joc.2025.04.08

After applying the K-means clustering algorithm to the principal component scores derived from the functional principal component analysis, the asset pool sample is classified into K distinct clusters.

3.2. Set trading mechanism

The pairs trading cycle is divided into two phases: the pairing phase and the trading phase. In this study, the first T=120 trading days serve as the pairing phase, followed by the (T+1) trading day as the trading phase. During the pairing phase, asset combinations exhibiting a Dependent Functional Cointegration relationship are identified; various parameters from the regression function are then estimated and subjected to dimensionality reduction, yielding hedging weight ratios for the trading phase. Subsequently, the model forecasts prices for the next trading day and computes the corresponding residual function series, serving as the basis for determining whether normal trading should proceed. This process is repeated until the final trading day, completing the pairs trading cycle.

Building on the adjustments made to the minimum-distance method, this study refines how deviations and reversion of spreads are measured and specifies how short and long positions are allocated for the paired assets. Accordingly, the entry and exit mechanisms are determined by the following trading workflow:

Open a position - choose to delay the opening of a position, that is, when the price difference breaks through the opening threshold and then returns to the historical threshold, to avoid the loss caused by the unilateral trend of price difference after triggering the opening conditions for the first time.

Close the position - select the spread to fall back to the close line and then close the position, add stop loss mechanism. When the price difference does not return to the mean value in time but keeps expanding, it is forced to close out the position of the stock held and exit to reduce the transaction risk.

Stop loss signal - If the threshold is triggered several times after the pairing portfolio has been placed throughout the trading period, clear the position and wait for the next trading signal. In the whole trading period, if the matched assets only show the signal of position establishment, but no signal of closing or stop loss, the asset portfolio will be forced to close to reduce the trading risk.

By making statistics on the dependent functional variables $Y_i^c(t_j)$ and $X_i^c(s)$ price sequences with different cointegration relations, the price difference sequence after the dependent functional cointegration regression can be obtained:

https://doi.org/10.7441/joc.2025.04.08

$$\begin{aligned} spread &= e_i(t_j) = Y_i^c(t_j) - \hat{Y}_i^c(t_j) \\ &= Y_i^c(t_j) - \int \hat{\beta}(s,t)\hat{X}_i^c(s)ds \end{aligned} \tag{20}$$

Where, $Y_i^c(t_j)$ represents the sample curve reconstructed by dependent function, and $\hat{Y}_i^c(t_j)$ represents the result of cointegration regression, estimation formula of average historical price difference μ, and standard deviation estimator σ, represented by:

$$\begin{aligned} \mu &= N^{-1}\sum_{i=1}^{N} e_i(t) \\ \sigma &= \sqrt{N^{-1}\sum_{i=1}^{N}(e_i(t)-\mu)^2} \end{aligned} \tag{21}$$

the trading signal in this paper is set to be traded again in the same trading day after opening and closing positions. The specific entry and exit mechanism is set as follows:

Short spread opening line: $Spread1 > \mu + \gamma\sigma$;

Long spread opening line: $Spread1 < \mu - \gamma\sigma$

Short spreadcloseout line: $Spread2 < \mu + \lambda\sigma$;

Long spread closeout line: $Spread2 > \mu - \lambda\sigma$;

3.3 Estimation of trading parameters

Caldeira and Moura (2013) set a threshold wherein a position is opened once the spread deviates by 2 standard deviations from its historical mean and remains open until the spread reverts to within 0.5 standard deviations or the trading period concludes. Nevertheless, for different datasets, parameter values can significantly affect both the timing and number of trades, ultimately influencing overall returns. Adopting a single, universal parameter setting may introduce statistical inaccuracies, as an excessively large opening threshold reduces trading opportunities, while an overly small threshold leads to more frequent trades, thereby increasing transaction costs and constraining the potential profit of each individual trade.

Accordingly, this study adopts a performance evaluation strategy that determines the optimal threshold values through an exhaustive search over historical data. Specifically, the best in-sample returns from the preceding 120 trading days serve as the assessment metric. With a step size of 0.1, the opening factor γ, closing factor λ, and stop-loss factor f are iteratively tested. These trading parameters are then re-estimated every 30 days. Concretely, the sample from trading days 1 to 120 is used to obtain the optimal γ, λ and *f*, which are applied to trading

https://doi.org/10.7441/joc.2025.04.08

days 121 to 150. Next, a new set of estimates is derived from the sample spanning trading days 31 to 151 to guide trades for days 151 to 180, and so forth.

Due to the market's high volatility and instability, historical data may fail to fully capture future uncertainties. Leung et al. (2001) point out that incorporating forecasting mechanisms into the trading portfolio can yield better performance in turbulent markets. Consequently, in estimating trading parameters during the pairing phase, this study employs the rolling regression coefficients derived from the Dependent Functional Data Analysis method to dynamically adjust the proportions of paired assets. The entry and exit rules for pairs trading are set as follows:

When the spread $Spread\hat{1}(t_j) > \hat{\mu}(t_j) + \hat{\gamma}\hat{\sigma}(t_j)$ exceeds a specified threshold at time t_{j+1}^1, a position is opened by shorting the Y asset portfolio and going long the $\hat{\beta}(t_{j+1}^1, t_{j+1}^1)$ X asset portfolio. Once the spread reverts to $\hat{\mu}(t_j) + \hat{\lambda}\hat{\sigma}(t_j)$, a designated level at time t_{j+1}^2, the position is closed by going long Y and shorting $\hat{\beta}(t_{j+1}^1, t_{j+1}^1)$ X. Conversely, if the spread falls below $Spread\hat{1}(t_{j'}) < \hat{\mu}(t_{j'}) + \hat{\gamma}\hat{\sigma}(t_{j'})$, a given threshold at time $t_{j'+1}^1$, a position is opened by going long Y and shorting $\hat{\beta}(t_{j'+1}^1, t_{j'+1}^1)$ X. After the spread returns to $\hat{\mu}(t_{j'}) + \hat{\lambda}\hat{\sigma}(t_{j'})$, the corresponding revert level at time $t_{j'+1}^2$, the position is closed by shorting Y and going long $\hat{\beta}(t_{j'+1}^1, t_{j'+1}^1)$ X.

The ratio of the hedging of the two pairs of assets in each trade can be used to obtain a trade-by-trade gain. Firstly, the binary regression coefficient $\beta(t)$ obtained by rolling regression is converted into a single dimensional coefficient function:

$$\beta(t) = \int_T \tilde{\beta}(t, s)ds, s \le t \tag{22}$$

Suppose the hedging ratio of two paired asset portfolios X: Y is β : 1. Over the time interval $[t_1, t_2]$, Once the residual sequence generated by regressing portfolio Y on portfolio X (under the functional financial framework) crosses the entry threshold for a long position, a trade is initiated. As the trade proceeds, when the residuals reach the closing threshold, that position is

https://doi.org/10.7441/joc.2025.04.08

closed. Ignoring transaction costs, the total rate of return for this trade during the period in question is calculated by summing up the accumulated profits.

$$In(\frac{P^s_{t_2}}{P^s_{t_1}}) - \sum_{m=1}^{M} \int_T \beta_m(s,t)ds * InIn(\frac{P^l_{t_2,m}}{P^l_{t_1,m}}) \quad (23)$$

4. RESULTS

4.1. Dependent functional data reconstruction

This study selects the constituent stocks of the HS 300 in 2020 as the sample data for pairs trading, utilizing intraday 5-minute closing price data sourced from the RESSET database. Specifically, we exclude ST stocks, remove stocks with more than two days of missing trading data, and discard trading days with more than five missing observations. For days with fewer than five missing observations, we apply the K-nearest neighbor (KNN) interpolation method to impute missing values. Ultimately, our refined dataset comprises 269 stocks, each containing data for 236 trading days. Trading hours span from 9:35 to 11:30 a.m. and 1:00 to 3:00 p.m. To mitigate potential intraday volatility biases arising from overnight and midday effects, the initial observation of each trading day is excluded in our empirical calculations.

Each stock's daily intraday closing price data are first smoothed using a roughness penalty approach, effectively converting discrete high-frequency observations into continuous functional data. Subsequently, we estimate the mean function and long-run covariance function, selecting the number of principal components based on a cumulative variance threshold of 95%, and then conduct dependent functional principal component analysis. Finally, utilizing the Karhunen–Loève (K-L) expansion, we reconstruct smooth and continuous functional curves from the original discrete price data. This functional reconstruction method not only efficiently achieves dimension reduction and noise filtering but also accurately captures the general market trend, thereby preserving essential informational content in high-frequency intraday data.

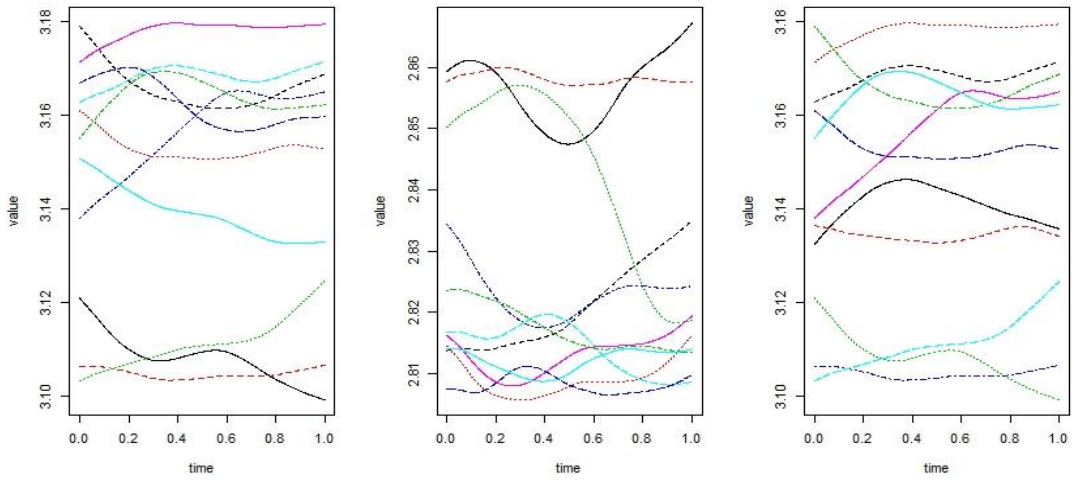


Figure 1. Current price fitting curve

https://doi.org/10.7441/joc.2025.04.08

We apply the K-means clustering method to classify all stocks within the selected stock pool, subsequently conducting pairs testing and trading within each cluster. Using the first 10 functional principal component scores derived from functional data reconstruction for the initial 100 trading days of each stock, we determine the optimal number of clusters by employing the Calinski–Harabasz (CH) index. According to the calculated CH index, the optimal number of clusters is determined to be 8, at which point the proportion of between-cluster sum of squares to total sum of squares reaches 96.10%, indicating excellent clustering performance. The resulting number of stocks in each cluster is as follows: 47, 35, 5, 43, 24, 48, 20, and 47, respectively.

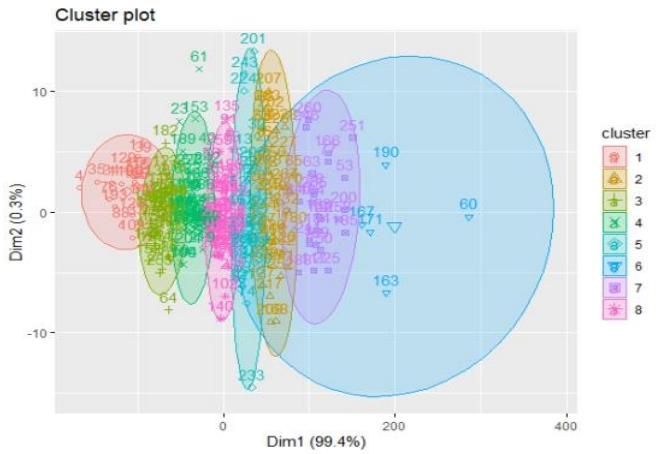


Figure 2. Results of K-means clustering analysis

We employ the stationarity testing procedure to examine the stationarity of price series, selecting parameters based on a cumulative variance explained threshold of 95% from principal components. Specifically, we set the truncation parameter J=500 and utilize Monte Carlo simulations with 1,000 repetitions to approximate the distribution matrix, thus determining the critical values T_N^* and $T_N^{0*}(d)$ for evaluating the presence of unit roots in functional time series. At the 1% significance level, the results indicate that the number of stocks meeting the criteria for pair selection within each of the eight clusters is 2, 3, 3, 2, 1, 4, 0, and 3, respectively. For all selected samples, the original functional time series were nonstationary but became stationary after first-order differencing.

Subsequently, we apply a fully functional linear cointegration regression model to price series integrated of the same order, based on the stationarity testing results. By substituting the

https://doi.org/10.7441/joc.2025.04.08

regression coefficient functions into the regression equation, we obtain the fitted values $\hat{Y}_i(t)$ and derive the residual functional sequences. As illustrated in Figure , the residual series clearly exhibit a pronounced mean-reverting trend.

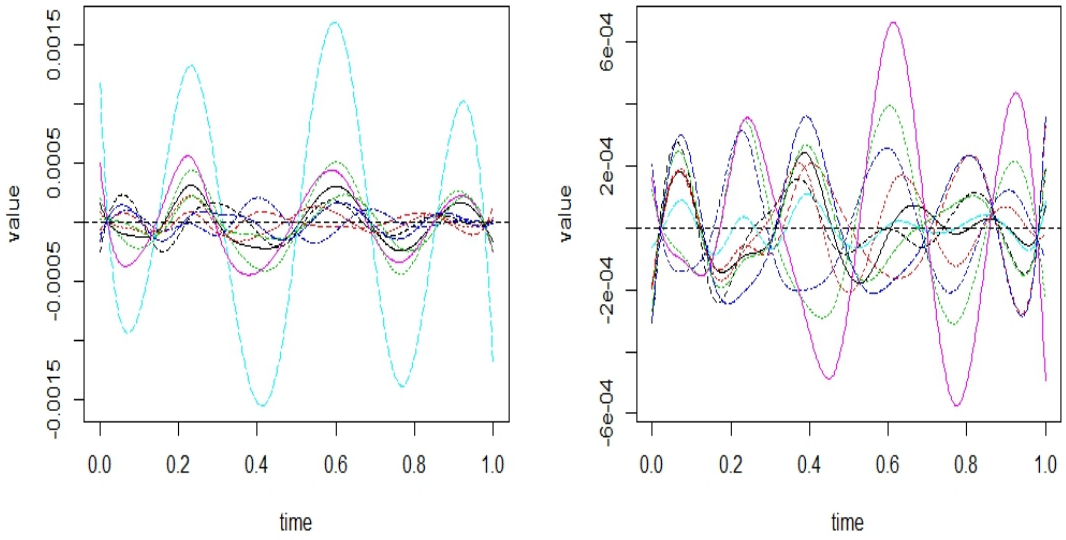


Figure 3. Residual sequences from the functional cointegration regression

We further conducted stationarity tests on the residual functional sequences obtained from the cointegration regression. The results indicate that within the eight clusters, there are respectively 2, 6, 6, 2, 7, 12, 0, and 6 pairs identified, totalling 34 stock pairs exhibiting significant cointegration relationships. These identified cointegrated pairs are thus eligible for subsequent pairs trading strategies.

Table1. Cointegration relationships of selected stock pairs

	Stock1		Stock2	
	T^N	T^0	T^N	T^0
C1	0.40	0.35	0.41	0.37
C2	0.45	0.36	0.44	0.36
	0.42	0.38	0.42	0.36
	0.41	0.39	0.43	0.41
C3	0.41	0.35	0.47	0.37

https://doi.org/10.7441/joc.2025.04.08

	0.44	0.40	0.44	0.39
	0.44	0.37	0.44	0.37
C4	0. 42	0. 38	0.43	0.36
C6	0.41	0.52	0.43	0.49
	0.42	0.47	0.29	0.36
	0.41	0.35	0.47	0.37
	0.27	0.35	0.44	0.40
	0.44	0.39	0.27	0.37
	0.44	0.37	0.44	0.37
C8	0.46	0.39	0.43	0.38
	0.45	0.36	0.43	0.38
	0.42	0.40	0.42	0.38

4.2. Profitability comparison

According to the cointegration test results, a total of 34 stock pairs exhibited significant functional cointegration relationships. To select higher-quality pairs for trading, this study further applied the correlation coefficient criterion, choosing stock pairs with both a functional cointegration relationship and an average correlation coefficient exceeding 0.95. Ultimately, 28 stock pairs were selected for the pairs trading strategy. To more intuitively demonstrate the specific performance of the proposed dependent functional cointegration pairs trading strategy, this study also constructed a traditional cointegration model as a control group. Both models shared identical transaction costs, trading fees, and trading rules. Their arbitrage performance was comprehensively compared from multiple perspectives, including trading parameter estimation, individual trade profitability, and cumulative returns over the entire trading period.

Follow the index trading rules, every five minutes can be traded on the futures market. In order to compare the statistical arbitrage results under different models, it is assumed that tradable non-integer stocks in the market, that is, the hedging ratio of X: Y of two paired asset portfolios is β : 1.If in $[t_1, t_2]$the spread is greater than the threshold, enter the market in the next t_{j+5} minute, sell Y asset and buy $\hat{\beta}(t_{j+5}, t_{j+5})$ X asset at the same time. When the spread returns, the original portfolio will be liquidated. Conversely, make the opposite operation for profit.

According to the test, the original price series of HS300 period based on the functional cointegration model and the traditional cointegration model are not stable, and the sample series is stable after the first-order difference. It is proved that the two sequences are first-order

https://doi.org/10.7441/joc.2025.04.08

integration, and the residual sequence of the regression equation has no unit root, so the cointegration relationship exists in both the current and future prices. Can carry on pair trade statistical arbitrage. In the whole matching trading process, 120 days were taken as the first matching period to estimate trading parameters, and trading parameters were re-estimated every 30 trading days. The following table shows the optimal threshold selection and returns of co-integration paired transactions in the sample of the quartic parameter estimation interval (the fifth parameter estimation interval less than 120 days is not considered).

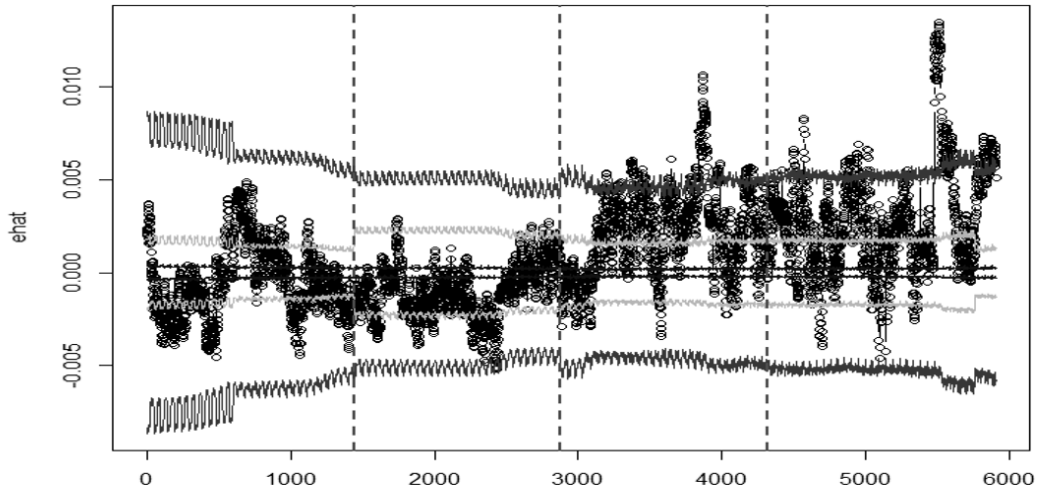
Table 2. Estimation of traditional cointegrating pair trading parameters

	1-120 days	31-150 days	60-181 days	90-211 days
stop loss *f*	2.4	2.1	2.4	2.2
open position γ	1	1	0.5	0.5
close out λ	0.1	0.1	0.1	0.1

Table 3. Estimation of dependent functional cointegration pairs trading parameter

	1-120 days	31-150 days	60-181 days	90-211 days
stop loss *f*	2.2	2	2.3	2.4
open position γ	0.5	0.9	0.8	0.8
close out λ	0.1	0.1	0.1	0.1

As indicated in the table, noticeable differences exist in the selected trading parameters across different estimation periods. This highlights that parameters used in pairs trading are, to some extent, influenced by the time-varying nature of market conditions.



https://doi.org/10.7441/joc.2025.04.08

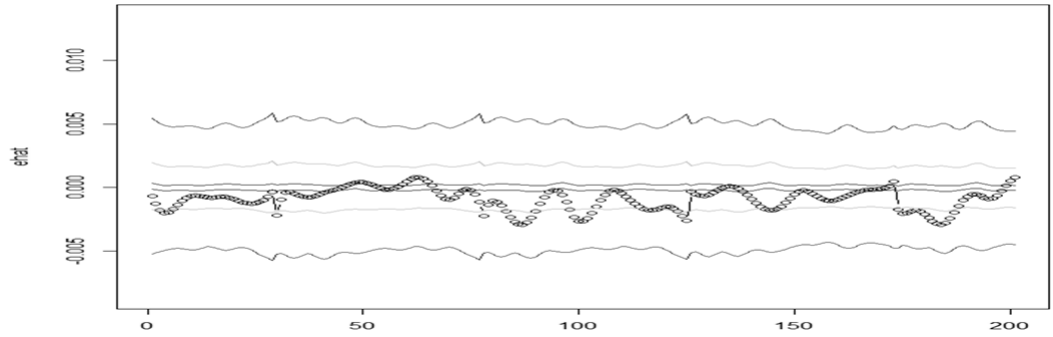


Figure 4. Pair transaction of dependent functional cointegration samples

The figure illustrates the trading performance of the dependent functional cointegration model over the entire trading period and at randomly selected 200 trading points. Specifically, the four intervals separated by dashed lines indicate periods during which trading parameters were re-estimated. The circled points represent the residual series used to determine whether trading thresholds were triggered. In the figure, the curves from top to bottom correspond to the stop-loss line, entry line, and exit line (close-position line) that trigger trading decisions, with the lower portion of the figure reflecting the opposite scenario.

When the circled points cross the entry line, a long position strategy is triggered—selling one stock and simultaneously buying the corresponding cointegrated stock pair. Subsequently, once the circled points revert to the exit line, the position is closed for profit-taking by implementing the reverse strategy. However, if the circled points move further upward or downward to reach the stop-loss line, positions are immediately closed to limit potential losses

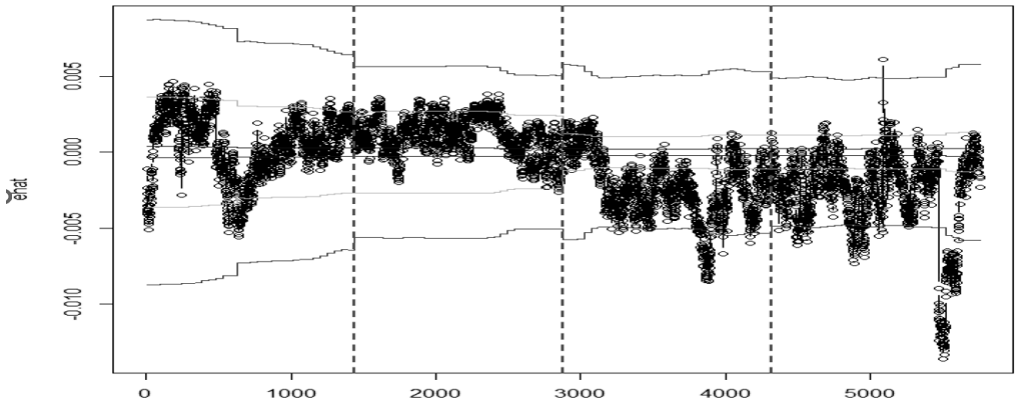


Figure 5. Sample matching transactions of traditional cointegration model

The residual sequences obtained from the functional cointegration regression clearly exhibit significant mean-reverting characteristics and higher trading frequency. In contrast, the

https://doi.org/10.7441/joc.2025.04.08

traditional cointegration pairs trading model demonstrates a notably lower frequency, resulting in fewer overall trading opportunities.

Table 5. Current arbitrage returns

	Traditional cointegration	Dependent functional cointegration
transaction times	7.14	14.78
average income	-2.34	16.54
max income	60.90	385.21
minimum income	-26.74	-192
Standard deviation	21.06	101.396
positive return ratio	61.56%	55.1%
Negative return ratio	38.43%	44.9%
Total revenue	34.357	1084.641

This paper evaluates pairs trading performance from three key perspectives: the average number of transactions per stock pair, the average profit per transaction, and the total accumulated profit for each pair over the entire trading period. A complete sequence from entry to exit is considered as a single transaction. The win rate per transaction is calculated as the ratio between the number of profitable transactions and the total number of completed transactions within each cluster. From the results, the dependent functional cointegration model shows an average holding period of 7–14 days per transaction, with an average profit of 16.54 per transaction, exhibiting right-skewed profitability distribution. Conversely, the traditional cointegration model generates fewer trades with longer average holding periods (17–24 days), significantly exceeding those of the dependent functional cointegration model.

Over the entire trading period, several observations emerge clearly: the dependent functional cointegration model yields the highest transaction frequency and the shortest average holding period, whereas the traditional cointegration model has the lowest transaction frequency and the longest holding periods. From the perspective of win rate, the traditional cointegration model achieves the highest win rate, making it more suitable for investors with lower risk tolerance, preferring a “low-risk, low-return” approach. In contrast, the dependent functional cointegration model demonstrates a more balanced distribution of positive and negative returns, thus aligning better with investors seeking “high-risk, high-return” strategies. Considering average returns, as well as maximum, minimum, and standard deviation of returns, both models generate positive average returns. However, substantial differences between the maximum and

https://doi.org/10.7441/joc.2025.04.08

minimum returns indicate significant volatility. The dependent functional cointegration model exhibits higher volatility, suggesting greater suitability for aggressive investors, whereas the traditional cointegration model, characterized by lower volatility and smaller standard deviation, is preferable for conservative investors. Given identical transaction costs and fees across the models, comparisons of total returns clearly demonstrate that the dependent functional cointegration model significantly outperforms the traditional model. Therefore, the dependent functional cointegration model proposed in this study, based on long-run covariance estimation, provides the most favorable outcomes and highest profitability in statistical arbitrage strategies.

5. CONCLUSIONS

High-frequency trading features high speed, high frequency, and extensive data processing capabilities, driving financial institutions, hedge funds, and other market participants to pursue continuous technological innovations in tools and methods, in order to compete for market pricing power and resource allocation. In this study, we construct a novel pairs trading strategy based on a dependent functional cointegration model to investigate trading and investment strategies in stock markets. While preserving data stochasticity, this model adopts a nonparametric framework, accurately capturing short-term mean reversion in prices and increasing trading frequency to enhance returns. Accordingly, it facilitates the creation of a universal quantitative trading system and provides a highly competitive trading method in high-frequency markets.

Specifically, during data preprocessing, this study accounts for the dependent and functional characteristics of high-frequency data. By using a long-run covariance estimator based on the untruncated Bartlett kernel, we correct the estimation bias inherent in functional data analysis under the assumption of independent and identically distributed data. Employing dependent functional data analysis, we reconstruct fitted functional curves from discrete trading prices and identify cointegrated asset pairs that are integrated of the same order. In the pairing stage, grounded in the core logic of pairs trading, mean reversion, and hedged trading methods, we refine existing minimum-distance and random-spread models to develop a dynamic hedging strategy, leveraging stationary spread series under mean reversion. We further set the trading cycle, entry and exit signal mechanisms, and trading parameters, before ultimately evaluating the cointegration-based pairs trading strategy with the goal of maximizing returns.

The results show that this improved dependent functional cointegration model not only surpasses the limitations of traditional statistical methods in processing high-frequency data, but also provides a more precise depiction of dependency and dynamic patterns in financial markets. Compared with conventional pairs trading approaches, it more effectively integrates

https://doi.org/10.7441/joc.2025.04.08

rolling regression coefficients to dynamically adjust hedge ratios between paired assets, offering traders a more reasonable entry-exit mechanism and asset allocation. In both in-sample and out-of-sample tests, it exhibits significantly superior returns.

Based on these findings, relevant policy recommendations are proposed. Trading institutions should strengthen data management and dynamically adjust models, establishing rational investment strategies for effective risk assessment and identifying profit opportunities. Meanwhile, regulatory agencies need to optimize market structures and enhance cross-market coordination. Looking ahead, this dependent functional cointegration-based pairs trading method could be extended to global markets and other asset classes, incorporating diverse trading rules and liquidity features to conduct more targeted research.

Funding:

1. Fujian Province Social Science Fund Project. “Research on financial asset dependent function volatility model and risk management.” FJ2024MGCA020.

2. Huaqiao University High-level Projects. “Research on the microstructure of stock index futures market based on the perspective of dependency function data analysis.” 23SKBS009.

REFERENCES

1. MacKinlay, A. C., & Ramaswamy, K. (1988). Index-futures arbitrage and the behavior of stock index future prices. *Review of Financial Studies, 1*(2), 137-158.

2. Gatev, E., Goetzmann, W., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies, 19*(3), 797-827.

3. Elliott, R., Hoek, V. D., & Malcolm, W. P. (2005). Pairs trading. *Quantitative Finance, 5*(3), 271-276.

4. Engle, R. F., & Granger, C. W. J. (1987). Cointegration and error-correction: Representation, estimation and testing. *Econometrica, 55*(2), 251-276.

5. Vidyamurthy, G. (2004). *Pairs trading: Quantitative methods and analysis*. John Wiley & Sons.

6. Huck, N., & Afawubo, K. (2015). Pairs trading and selection methods: Is cointegration superior? *Applied Economics, 47*(6), 599-613.

7. Rad, H., Low, R. K. Y., & Faff, R. (2016). The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance, 16*(10), 1541-1558.

https://doi.org/10.7441/joc.2025.04.08

8. Blázquez, M. C., & Román, C. P. (2018). Pairs trading techniques: An empirical contrast. *European Research on Management and Business Economics, 24*(3), 160-167.

9. Vychytilová, J. (2015). Linkages among U.S. Treasury bond yields, commodity futures and stock market implied volatility: New nonparametric evidence. *Journal of Competitiveness, 7*(3), 143-158.

10. Huang, Z., & Martin, F. (2019). Pairs trading strategies in a cointegration framework: Back-tested on CFD and optimized by profit factor. *Applied Economics, 51*(22), 2436-2452.

11. Madden, S. (2012). From databases to big data. *IEEE Internet Computing, 16*(3), 4-6.

12. Leung, M. T., Daouk, H., & Chen, A. (2001). Using investment portfolio return to combine forecasts: A multiobjective approach. *European Journal of Operational Research, 134*(1), 84-102.

13. Ramsay, J. O. (1982). When the data are functions. *Psychometrika, 47*(4), 379-396.

14. Ramsay, J. O., & Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological), 53*(3), 539-561.

15. Ramsay, J. O., & Silverman, B. W. (2002). *Applied functional data analysis: Methods and case studies* (2nd ed.). Springer.

16. Huang, S. P., Quek, S. T., & Phoon, K. K. (2001). Convergence study of the truncated Karhunen-Loeve expansion for simulation of stochastic processes. *International Journal for Numerical Methods in Engineering, 52*(9), 1029-1043.

17. Müller, H. G., Sen, R., & Stadtmüller, U. (2011). Functional data analysis for volatility. *Journal of Econometrics, 165*(2), 233-245.

18. Dauxois, J., Pousse, A., & Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis, 12*(1), 136-154.

19. Boubaker, S., Liu, Z., Lu, S., & Zhang, Y. (2021). Trading signal, functional data analysis and time series momentum. *Finance Research Letters, 42*, 101933.

20. Tsay, R. S. (2016). Some methods for analyzing big dependent data. *Journal of Business & Economic Statistics, 34*(4), 673-688.

21. Das, S., Demirer, R., Gupta, R., & Mangisa, S. (2019). The effect of global crises on stock market correlations: Evidence from scalar regressions via functional data analysis. *Structural Change and Economic Dynamics, 50*, 132-147.

https://doi.org/10.7441/joc.2025.04.08

22. Hörmann, S., & Kokoszka, P. (2010). Weakly dependent functional data. *Annals of Statistics, 38*(3), 1845-1884.

23. Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications*. Springer.

24. Kokoszka, P., & Young, G. (2017). Testing trend stationarity of functional time series with application to yield and daily price curves. *Statistics and Its Interface, 10*(1), 81-92.

25. Horváth, L., Kokoszka, P., & Rice, G. (2014). Testing stationarity of functional time series. *Journal of Econometrics, 179*(1), 66-82.

26. Caldeira, J. F., & Moura, G. V. (2013). Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy. *Brazilian Review of Finance, 11*(1), 49-80.

Contact information

Danni Wang, Ph.D.
Huaqiao University
Department of Economics and Finance
China
E-mail: wangdanni@hqu.edu.cn

Zhifang Su, Ph.D.
Huaqiao University
Department of Economics and Finance
China
E-mail: suzufine@hqu.edu.cn

https://doi.org/10.7441/joc.2025.04.08