



Improving Cointegration-Based Pairs Trading Strategy with Asymptotic Analyses and Convergence Rate Filters

Yen-Wu Ti¹ · Tian-Shyr Dai^{2,3} · Kuan-Lun Wang⁴ · Hao-Han Chang² · You-Jia Sun²

Accepted: 14 December 2023 / Published online: 19 January 2024
© The Author(s) 2024

Abstract

A pairs trading strategy (PTS) constructs a mean-reverting portfolio whose logarithmic value moves back and forth around a mean price level. It makes profits by longing (or shorting) the portfolio when it is underpriced (overpriced) and closing the portfolio when its value converges to the mean price level. The cointegration-based PTS literature uses the historical sample mean and variance to establish their open/close thresholds, which results in bias thresholds and less converged trades. We derive the asymptotic mean around which the portfolio value oscillates. Revised open/close thresholds determined by our asymptotic mean and standard derivations significantly improve PTS performance. The derivations of asymptotic means can be extended to construct a convergence rate filter mechanism to remove stock pairs that are unlikely to be profitable from trading to further reduce trading risks. Moreover, the PTS literature oversimplifies the joint problem of examining a stock pair's cointegration property and selecting the fittest vector error correction model (VECM). We propose a two-step model selection procedure to determine the cointegration rank and the fittest VECM via the trace and likelihood ratio tests. We also derive an approximate simple integral trading volume ratio to meet no-odd-lot trading constraints. Experiments from Yuanta/P-shares Taiwan Top 50 Exchange Traded Fund and Yuanta/P-shares Taiwan Mid-Cap 100 Exchange Traded Fund constituent stock tick-by-tick backtesting during 2015–2018 show remarkable improvements by adopting our approaches.

Keywords Pairs trading · Asymptotic mean-reverting level · Convergence rate filter

Abbreviations

PTS	Pairs trading strategy
VECM	Vector error correction model
DM	Distance method
ECM	Error correction model
ADF	Augmented Dickey–Fuller

Extended author information available on the last page of the article

VAR	Vector autoregressive
AR	Autoregressive model
BIC	Bayesian information criterion
LRT	Likelihood ratio test
PNL	Profit and loss
ASV	Asymptotic mean
AVG	Average
MDD	Maximum draw down
SAM	Sample mean
T-S	Two-step model selection
M3	Model 3

1 Introduction

The pairs trading strategy (PTS), a well-known market-neutral strategy that originated on Wall Street in the 1980s, has become increasingly prevalent, as it is difficult to determine whether a stock is over or underpriced (Chang et al., 2020). Instead of guessing at unpredictable market tendencies, the PTS forms a portfolio by longing one stock and shorting another at a specific ratio to eliminate this risk. The specific linear combination of logarithmic values of these two stocks, referred to as the "spread," moves back and forth around a particular level and thus exhibits mean reversion. When the spread process deviates excessively, an investor can buy the undervalued stock and short the overvalued stock; the investor can close the portfolio to make a profit when the spread reverts to its mean-reverting level.

Among the various techniques used to construct a PTS suggested in Krauss (2017), our paper focuses on the Johansen cointegration approach, which is shown to outperform other methods, as described in Do and Faff (2010, 2012), Huck and Afawubo (2015), and Rad et al. (2016). Engle and Granger (1987) represent the cointegration relationship by a vector error correction model (VECM), and Vidyamurthy (2004) presents the theoretical framework of cointegration-based PTS in terms of a VECM. Although a PTS can also be implemented by the Engle and Granger cointegration test, as stated in Lin et al. (2006), this test is characterized by the asymmetry problem, which argues which stock price should be treated as either the explanatory or the response variable. This paper proposes an improved method based on the Johansen cointegration test to calibrate the price processes of a pair of assets eligible for a PTS, and we conduct comprehensive experiments to demonstrate the superiority of our approach.

Appropriate selection of proper VECM models has been studied by Johansen (1995). However, the cointegrated-based PTS literature either selects models heuristically (see Li et al. (2014) and Rad et al. (2016)) or even ignores this issue. In addition, we find that Johansen's model selection begins from the narrowest VECM and may excessively eliminate eligible stock pairs for the PTS that belong to more comprehensive models. To address this issue, we will propose a two-step model selection method that employs the trace test in the first stage to calibrate the logarithmic stock price processes of each stock pair to identify eligible VECMs possessing

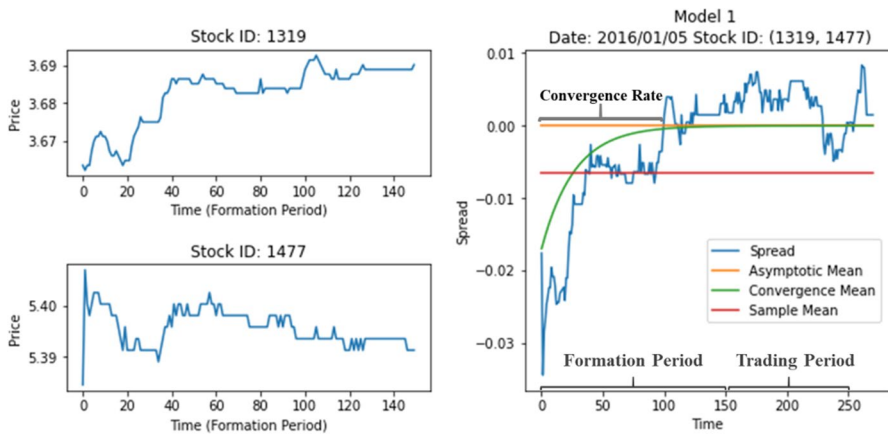


Fig. 1 Sample Illustration of Stock Price Processes, Spread Process, Sample Mean, Asymptotic Mean, and Convergence Mean. The spread process is constructed by a linear combination of intra-day price processes of stocks with tick symbols 1319 and 1477 on Jan 5th, 2016. The asymptotic mean is constructed by calibrating Model 1 defined in Sect. 3.3 with parameters $\alpha = (-0.02, 0.02)'$, $\beta = (1, -0.68)'$, and $M=0$ as derived in Sect. 3.2. The convergence rate measures how fast the convergence mean converges to the asymptotic mean described in Sect. 3.4. (Color figure online)

cointegration properties. Then, we select the fittest model from the candidates identified in the first stage with the likelihood ratio test in the second stage.¹ Experiments suggest that adopting our model selection significantly improves trading performance.

Vidyamurthy (2004) uses the sample mean and standard deviation of a spread to estimate the mean-reverting level and the open threshold for a PTS. This approach is widely adopted in the PTS literature, such as Huck and Afawubo (2015), Huck (2010), and Rad et al. (2016). Rad et al. (2016) find that a large portion of spread processes fail to converge to sample means,² which significantly erodes profits. Indeed, we also discover that the intraday spread process generally oscillates around the asymptotic mean (derived in Sect. 3.2) rather than the sample mean, as illustrated in Fig. 1. This is because heavy stock price fluctuations and high trading volumes typically occur at the beginning of each trading day, reflecting the impact of overnight information such as breaking news during market closed periods, as suggested in Tian (2007). Therefore, these price fluctuations significantly bias sample means and standard deviations of spread processes, frequently resulting in the above-mentioned failure to converge. By rewriting a VECM with higher lag orders into a companion form with lag order 1 as in Hansen (2000), this paper explores the asymptotic behavior of the spread mean (as illustrated by the green line in Fig. 1). Then we derive the asymptotic limit of the spread mean (the orange line) and its standard

¹ Unlike our two-step method, which is based on statistical tests, some criterion-based methods address selection of VECMs with cointegration properties, such as Villani (2005).

² They compare the portion and returns of unconverged trades to those of converged trades in Table 5 of their study.

deviation by taking advantage of the recursive relation of the VECM selected by the aforementioned two-step method. Numerical results suggest that determining the mean-reverting level and the open threshold with our asymptotic results outperforms the naive adoption of sample means and standard deviations.

Indeed, rapid market change may break the cointegration relationship due to the occurrence of structural breaks, which results in significant loss (Huang et al. 2020). Sarmiento and Horta (2020) and Lu et al. (2022) reduce PTS trading risks by using machine learning models to filter out unprofitable stock pairs. However, these black-box models are not interpretable. To avoid the risk of overnight structural breaks occurring during the market closed period, we follow Stübinger (2019), Endres and Stübinger (2019), and Liu et al. (2017) by considering a high-frequency intraday PTS that limits the duration of our PTS to within a day. Costs are also suppressed since the Taiwan Stock Exchange provides a 50% discount on transaction costs on intra-day trading. To address structural breaks that occur during trading, we take advantage of the derivation of the asymptotic mean to measure the convergence rate of the spread (proxied by the convergence mean in Fig. 1) to its mean-reverting level (denoted by the asymptotic mean); this rate reflects the likelihood for a spread to return to the asymptotic mean and make a profit. Thus, we filter less profitable stock pairs by removing corresponding spreads with low convergence rates to improve win rates and the Sharpe ratio, as confirmed in our backtesting. Note that the ratio for shorting one stock and longing another derived by calibrating VECM parameters could be a non-simple irreducible ratio.³ In practice, however, the number of traded stocks cannot be infinitely divided, and intraday trading of odd lots is not allowed. Arbitrarily amplifying this trading ratio would erode PTS profits since a large trading volume could significantly influence stock prices and cause slippage. To our knowledge, this PTS study is the first to address this problem by adjusting the ratio to a nearby simple integer ratio with small magnitudes; thus, we reasonably trade indivisible stocks under the constraint of limited liquidity. To prevent such adjustment from weakening the stationary property of the spread and hence deteriorating PTS performance, we remove stock pairs whose adjusted spread processes fail to reject the null hypothesis (the presence of unit roots) of the augmented Dickey–Fuller (ADF) test. Experiments show that these proposed adjustments also yield profitable results.

The rest of this paper is organized as follows. Section 2 details the required background knowledge of PTS. Section 3 describes the derivations of asymptotic means and standard derivation used to build the open/close thresholds. These derivations can be further extended to remove spread processes with low convergence rates to reduce trading risks. Then we describe the two-step model selection method and adjustments to the trading ratio to meet the no-odd-lot and liquidity constraints without significantly sacrificing profitability. In Sect. 4, to confirm the superiority of our model, we analyze the tick-by-tick backtesting on the 0050.TW and 0051.TW

³ For example, a cointegrating vector $(1, -2.1)$ with latest stock prices $(100, 50)$ denotes the ratios of trading shares are $(\frac{1}{100}, -\frac{2.1}{50})$.

constituent stocks, which were the top companies with the largest market capitalization from 2015 to 2018.⁴ We conclude in Sect. 5.

2 Preliminaries

In this section, we first survey the relevant PTS literature. Then we introduce the key properties of a vector autoregressive (VAR) model and a VECM required to implement our PTS improvements, after which we express the cointegration property under the VECM model.

2.1 Literature Review

Statistical and time series analyses are increasingly used in economic and financial applications. For instance, Xu and Ye (2023) optimize investments using value at risk measures, whereas Tsionas et al. (2023) apply Bayesian methods to examine decision making under uncertainty. Maiti et al. (2023) forecast India's post-COVID GDP with a factor-augmented error correction model, Liu et al. (2023) predict stock returns using a vector-based clustering ensemble, He et al. (2023) implement online portfolio selection with an aggregating algorithm, and Sapna and Mohan (2023) predict the implied volatility of Ethereum options. Bampinas and Panagiotidis (2015, 2016) explore the use of cointegration to hedge against inflation with gold, silver, and stock prices. This paper focuses on the application of statistical and time series analyses in pairs trading. We follow Krauss (2017) in dividing and surveying PTS techniques according to the five following groups. First, the distance method (abbreviated as DM) adopted in Bowen (2010), Do and Faff (2010, 2012), and Gatev et al. (2006) calculates the Euclidean distance of two stock price processes to measure their co-movements; they then apply PTS on stock pairs with close co-movements. Second, Engle and Granger (1987) and Johansen (1988) develop a cointegration test to determine the existence of a linear combination of stock prices that makes a stock portfolio's value a stationary process. Since the mean of the stationary process does not change with time, an investor can buy (sell) the portfolio when its value is below (above) the mean and close the position to make a profit when the value converges back. Vidyamurthy (2004), Lin et al. (2006), Li et al. (2014), Rad et al. (2016), Malkidis and Fountas (2020), and Lin and Tan (2023) apply the aforementioned cointegration tests to construct PTS-eligible stock pairs. Yan et al. (2020) find that the path-dependent cointegration model can further improve PTS performance. Third, Elliott et al. (2005) and Cummins and Bucca (2012) find optimal trading rules for PTS-eligible stock pairs under different time series models, such as the autoregressive moving average model, the generalized autoregressive conditional

⁴ Most studies execute overnight PTS by setting a long inter-day trading period and making trading decisions based on every day's close price. Thus a sufficiently long backtesting period is required to examine their PTS models' performance. Since trading decisions are based on intraday tick data, trading can be performed during one trading day. Experimental results show that this four-year backtesting period generates around 30,000 trading opportunities for analyses. Note that intraday tick data are not easily collected as the daily close price data, which might explain why few PTS studies address intraday trading. We have made every effort to gather tick data spanning four years for our experiments.

heteroscedasticity model, and the Ornstein–Uhlenbeck model. On the same premises, the fourth approach uses stochastic control theory to determine the best asset allocations for a PTS (see Do and Faff (2010)). The fifth category denotes other techniques, such as the principal component analysis method proposed by Avellaneda and Lee (2010), Krauss and Stübinger (2017), and Ferreira (2008), who reduce the dimension of the stock return vectors to find co-moving stock pairs. In addition, the copula method analyzed in Liew and Wu (2013), Rad et al. (2016), and Stander et al. (2013) estimates the joint distribution of the time series of stock returns to determine the overvalue or undervalue likelihoods, whereas Huck (2009) and Huck (2010) use multi-criteria decision techniques to rank and detect under- and over-valued stocks for PTS. This paper focuses on improving PTS construction based on the Johansen cointegration test, which outperforms other methods as studied in the following literature. Specifically, Do and Faff (2010, 2012) find that the DM performance declines over time and is generally not profitable when transaction costs are taken into account. Huck and Afawubo (2015) perform large-scale empirical tests on S&P 500 to show that the cointegration approach outperforms DM. Rad et al. (2016) also show that the cointegration approach exhibits better after-transaction-cost performance than the copula method on the entire US equity market from 1962 to 2014.

There are three different models (excluding trend-stationary models) with different determinist terms studied in Johansen (1995). A proper cointegration model selection method could more feasibly capture the characteristics of stock price processes and hence improve the PTS performance. However, the model selection issue is simplified or overlooked in the cointegration-based PTS literature. For example, Li et al. (2014) directly select the most generalized type (Model 3 defined in Sect. 3.3). Rad et al. (2016) directly use the simplest type (Model 1). Other papers even ignore this issue. On the other hand, Johansen (1995) suggests that the conventional method of model selection focuses on the existence of deterministic terms and begins the trace test from the narrowest model, Model 1, and then extends to broader models. However, this approach may excessively eliminate eligible stock pairs for PTS that belong to broader models. To mitigate this issue, we apply the trace test from the widest model, Model 3, and then extend this to Model 1. Additionally, if the price process of a stock pair cannot be rejected from belonging to more than one model, we enhance the model fitness by employing the likelihood ratio test, as discussed in Sect. 3.3. Additionally, Otero et al. (2022) demonstrate that the power of the Johansen cointegration test is influenced by the duration of the sample period and the number of observations. We align our study with this finding by adjusting both the length and frequency of the sample period. Subsequently, we evaluate the test's effectiveness in terms of trading performance.

Much of the literature improves PTS performance using machine learning techniques, such as Kim and Kim (2019), Sarmiento and Horta (2020), Kuo et al. (2021), Lu et al. (2022), and Kuo et al. (2022). Although these approaches involve the use of deep or reinforcement learning techniques to determine optimal open/stop-loss thresholds or detect structural breaks, statistical methods such as the cointegration approach are still needed to identify stock pairs that are eligible for the PTS and to determine feasible investment capital ratios and mean-reverting levels. Thus, improving the quality of cointegrated-based PTS is important.

2.2 Required Properties for VAR and VECM

Başcı and Karaca (2013) suggest that a multivariate time series can be modeled by a VAR model, a multivariate generalization of an autoregressive model (AR). Let $I(d)$ denote that the order of integration of a time series is d . A k -dimensional $I(1)$ process X_t following a VAR model of order p , or VAR(p) for short, can be expressed as

$$X_t = c_t + \prod_1 X_{t-1} + \dots + \prod_p X_{t-p} + u_t, t = 1, \dots, T \text{ and } p > 0 \quad (1)$$

where c_t is a k -dimensional deterministic vector, Π_i is a $k \times k$ matrix for $1 \leq i \leq p$, and u_t is a k -dimensional independent and identically distributed Gaussian white noise with mean zero and a positive definite covariance matrix Σ , i.e., $u_t \stackrel{\text{iid}}{\sim} \text{MVND}(0, \Sigma)$. The VAR(p) process in Eq. (1) can be rewritten in terms of a VECM as

$$\Delta X_t = c_t + \Pi X_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + u_t = c_t + \alpha \beta' X_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + u_t \quad (2)$$

where $\Delta X_t = X_t - X_{t-1}$, $\Pi = -I + \sum_{i=1}^p \Pi_i$, and $\Gamma_i = -\sum_{j=i+1}^p \Pi_j$. If the rank of Π is a constant r larger than 0 and there exist two $k \times r$ matrices α and β such that $\Pi = \alpha \beta'$ and $\text{Rank}(\alpha) = \text{Rank}(\beta) = r$, then X_t is cointegrated with r linearly independent cointegrating vectors, as stated in Theorem 4.2 of Johansen (1995). Based on the above, Hansen (2000) deduces the following lemma about the companion form of matrices.

Lemma 1. Let $\tilde{X}_t = (X_t', X_{t-1}', \dots, X_{t-p+1}')'$ and the orthogonal complement of a matrix M be M_\perp . Then VECM(p) defined in Eq. (2) can be transferred to a companion form.

$$\begin{aligned} \Delta \tilde{X}_t &= \tilde{\Pi} \tilde{X}_{t-1} + \tilde{c}_t + \tilde{u}_t \\ &= \tilde{\alpha} \tilde{\beta}' \tilde{X}_{t-1} + \tilde{c}_t + \tilde{u}_t \end{aligned}$$

where

$$\begin{aligned} \Delta \tilde{X}_t &= (\Delta X_t', \Delta X_{t-1}', \dots, \Delta X_{t-p+1}')' \\ \tilde{\Pi} &= \begin{bmatrix} \alpha \beta' + \Gamma_1 & \Gamma_2 - \Gamma_1 & \Gamma_3 - \Gamma_2 & \dots & \Gamma_{p-1} - \Gamma_{p-2} & -\Gamma_{p-1} \\ I & -I & 0 & \dots & 0 & 0 \\ 0 & I & -I & & & \\ 0 & 0 & I & \ddots & & \\ \vdots & \vdots & & & & \\ 0 & 0 & & & -I & 0 \\ & & & & I & -I \end{bmatrix} \end{aligned}$$

$$\tilde{\alpha} = \begin{bmatrix} \alpha & \Gamma_1 & \dots & \Gamma_{p-1} \\ 0 & I & & 0 \\ \vdots & & \ddots & \\ 0 & & & I \end{bmatrix}$$

$$\tilde{\beta} = \begin{bmatrix} \beta & I & 0 & \dots & 0 \\ 0 & -I & I & & \\ \vdots & & -I & \ddots & \\ 0 & & & \ddots & I \\ & & & & -I \end{bmatrix}$$

$$\tilde{c}_t = (c'_t, 0_{\{p-1\}})'$$

$$\tilde{u}_t = (u'_t, 0_{\{p-1\}})'$$

$$\tilde{\alpha}_\perp = \begin{bmatrix} \alpha_\perp \\ -\Gamma'_1 \alpha_\perp \\ \vdots \\ -\Gamma'_{p-1} \alpha_\perp \end{bmatrix}$$

and

$$\tilde{\beta}_\perp = \begin{bmatrix} \beta_\perp \\ \beta_\perp \\ \vdots \\ \beta_\perp \end{bmatrix}$$

2.3 Cointegrated VECM for PTS

The cointegration property of a collection of time series with integration order m denotes that the existence of a linear combination of series from this collection has an integration order less than m . This useful property can be used to construct a linear combination of log-stock prices, referred to as the "spread," following a stationary process (i.e., an $I(0)$ process) given that the log-stock price processes are non-stationary. The mean-reverting property of the stationary process allows us to apply a PTS to the stock portfolio to make profits. For convenience, define two logarithmic processes $\{X_{1t}\}$ and $\{X_{2t}\}$ by $X_{1t} = \ln(X_{1,t})$ and $X_{2t} = \ln(X_{2,t})$, where $\{X_{1,t}\}$ and $\{X_{2,t}\}$ denote the price processes of two arbitrary stocks. The stock portfolio is constructed by applying the Johansen cointegration test on the vector of time $X_t \equiv (X_{1t}, X_{2t})$. The cointegration property entails that linear combination(s) $\beta' X_t$ are stationary processes, where each column of β (defined in Eq. (2)) denotes a cointegrating vector. Here we combine the proof results in Hansen (2000) that will be used later as follows.

Lemma 2. (Hansen, 2000) *Let $\Gamma = I - \sum_{i=1}^{p-1} \Gamma_i \cdot \left| \alpha'_\perp \Gamma \beta_\perp \right| \neq 0$ if and only if the absolute value of the eigenvalues of $\left(I + \tilde{\beta}' \tilde{\alpha} \right)$ are less than 1 for any VECM model.*

In our backtest experiments, to ensure the $\left| \alpha'_\perp \Gamma \beta_\perp \right| \neq 0$ property holds in our PTS model, we calibrate the price processes of each stock pair by Eq. (2) and exclude stock pairs whose estimated determinants of $\alpha'_\perp \Gamma \beta_\perp$ are sufficiently small.

The property of "less-than-one" eigenvalues entails that $\left(I + \tilde{\beta}' \tilde{\alpha} \right)^t$ decays exponentially with increments in t . This result is useful to derive the asymptotic mean and variance of the spread process $\beta' X_t$ as in Sect. 3.

3 Model Development

This section begins by detailing how we preprocess the transaction data and by outlining the structure of our PTS in Sect. 3.1. Following this, we highlight our major contributions towards enhancing PTS performance. The sample mean/variance used to build trading thresholds in the literature is replaced by the asymptotic mean/variance derived in Sect. 3.2. Section 3.3 selects the fittest VECM models that possess the cointegration property based on the two-step method which makes use of the trace test and the likelihood ratio test. Section 3.4 proposes the convergence rate filter based on the asymptotic mean/variance analyzed in Sect. 3.2 to filter out stock pairs that are unlikely to be profitable. Section 3.5 converts the non-integer trading ratio from the cointegrating vector into the approximated irreducible integer ratio to reasonably trade indivisible stocks with limited liquidity or to account for no-odd-lot trading limitations. Finally, Sect. 3.6 describes how we construct trading thresholds (i.e., open and close thresholds) and the execution process of the PTS.

3.1 Data Preprocessing and Structure of the PTS

Given that transaction expenses and price slippage heavily influence PTS profitability, in our later experiments we conduct intraday pairs trading using the constituent stocks of the 0050 TT YUANTA/P-SHRS TW and 0051 TT YUANTA/P-SHRS TW MID-CAP 100 ETFs listed on the Taiwan Stock Exchange without holding positions overnight. This is because these stocks have better liquidity (i.e., lower price slippage), and the Taiwan Stock Exchange grants 50% discounts for intraday trading.⁵ This reduction of slippage and cost significantly increases both win rates and profits. In addition, intraday trading also prevents investment losses caused by the ineffectiveness of the cointegration property due to changing market conditions and breaking news during market closed periods. On each trading day, the Taiwan Stock Exchange begins at 9:00 a.m. and closes at 1:30 p.m.; we use the first 150 min

⁵ The transaction cost is 0.3% for ordinary cross-day trading but is reduced to 0.15% for intraday trading.

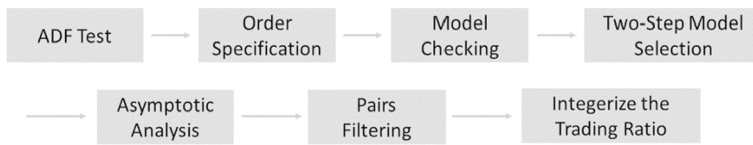


Fig. 2 Flowchart for Finding Eligible Stock Pairs and Constructing Trading Strategies

(i.e., the formation period) to find PTS-eligible stock pairs with the cointegration test and trade these pairs during the rest of the day. The t -th minute average stock price of the i -th stock, $X_{i,t}$, during the formation period is calculated as the volume-weighted average stock price that occurs at that minute and the log stock price series $\{X_{it}\}$ with length 150 is constructed by taking the logarithm on the minute-average price of the i -th stock. Otero et al. (2022) find that the power of a cointegration test is affected by the duration of the sample period—here referred to as the formation period—and the number of observations. To test their findings, in our later experiments we reduce the formation period from 150 to 100 min and change the sampling frequencies to assess the impact on PTS performance.

After this data preprocessing, we construct PTS-eligible stock pairs as illustrated in Fig. 2. We first filter all $\{X_{it}\}$ without unit root(s) via the ADF test proposed in Kwiatkowski et al. (1992) to ensure that all the log-stock price processes $\{X_{it}\}$ adopted for PTS are non-stationary⁶ (i.e., the ADF test stage). Next, we pick two different log-stock price processes to form a log-stock-price vector $X_t \equiv (X_{1t}, X_{2t})$ and then calibrate it with the VAR defined in Eq. (2).⁷ The order of the VAR, denoted as p , under a given deterministic term c_t , can be estimated by the Bayesian information criterion (BIC), as described in Stoica and Selen (2004) (i.e., the order specification stage in Fig. 2). The BIC is formulated as $\ln(\text{Max}(\Sigma, p)) + \frac{\ln(T)}{T}pk^2$, where $\text{Max}(\Sigma, p)$ denotes the determinant of the maximum likelihood estimate of Σ for a VAR(p) model, T denotes the number of samples, and k denotes the dimension of the VAR model as introduced in Sect. 2.2. To check whether the VAR(p) feasibly models the log-stock-price vector X_t , the third step estimates the residual vector a_t and then applies the Jarque–Bera test and the portmanteau test to test the normality and the autocorrelation of u_t (i.e., the model-checking stage).

To select the most suitable VECM model among those studied in Johansen (1995), we propose a two-step model selection process in Sect. 3.3. This process involves simultaneously determining eligible stock pairs for PTS and their corresponding models (i.e., the two-step model selection stage). We then use the selected model and its parameters to derive the asymptotic mean and variance for each stationary process (i.e., the asymptotic analysis stage), as detailed in Sect. 3.2. Next, we

⁶ The buy low, sell high strategy can be directly applied on a stationary stock price process without forming a cointegrated stock pair.

⁷ Indeed, the Johansen test can extract more cointegration relations from a larger number of stock price processes. However, our experiments find that the mean-reverting profit is significantly eroded by simultaneously trading more stocks. Thus, our paper focuses on the “two-stock” pair—or the univariate pairs discussed in Krauss (2017).

develop a convergence rate filter to remove unlikely profitable stock pairs that have low convergence rates (the pair filtering stage), as described in Sect. 3.4. The final stage converts the non-integer trading ratio of the stock pair into an integer one as in Sect. 3.5 to meet the no-odd-lot trading constraint.

Having completed the PTS settings using these procedures, we then calculate the trading thresholds (i.e., open and close thresholds) based on the asymptotic mean and variance. These thresholds are used to execute trades for each stock pair using the stock price tick data during the trading period, as discussed in Sect. 3.6.

3.2 Deriving Asymptotic Convergence Mean/Variance for the Spread

Based on the discussion in Sect. 2, we derive the values of the asymptotic mean and variance of the spread sequence $\beta'X_t$ used to determine the thresholds for opening and closing the portfolio in our PTS by the following theorems and lemmas. First, we introduce Lemma 3 derived by Hansen (2000) to analyze the convergence property of VECM(p). To evaluate the asymptotic convergent mean and the variance of $\beta'X_t$ with good readability, we first rewrite the stationary proofs of $\beta'X_t$ in Theorem 4.2 of Johansen (1995) with our notation in the former part of Theorem 1. The rewritten formulae are used to derive the mean/variance and to define the convergence rate filter in Sect. 3.4.

Lemma 3. (Hansen, 2000) *If $|\alpha'_\perp \Gamma \beta_\perp| \neq 0$, then $\tilde{\alpha}$ and $\tilde{\beta}$ defined in Lemma 1 have the following properties: For the eigenvalue λ of the matrix $I + \tilde{\beta}'\tilde{\alpha}$, $|\lambda| < 1$, the matrix $I + \tilde{\beta}'\tilde{\alpha}$ has no simple root and $(I + \tilde{\beta}'\tilde{\alpha})^t$ converges exponentially to 0 as t increases.*

The property $|\alpha'_\perp \Gamma \beta_\perp| \neq 0$ in our experiment is ensured by excluding stock pairs whose estimated determinants of $\alpha'_\perp \Gamma \beta_\perp$ are sufficiently small when calibrating the price processes of each stock pair by Eq. (2).

Theorem 1. *For any VECM(p) which is written as.*

$$\Delta X_t = \prod X_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + c_t + u_t$$

we assume $|\alpha'_\perp \Gamma \beta_\perp| \neq 0$ and rewrite the stationary proof of $\beta'X_t$ proposed in Johansen (1995) using our notation to improve readability. Next, we show that the conditional mean $E(\beta'X_t | X_0, X_{-1}, \dots, X_{-p+1})$ converges to a constant $J \left((-\tilde{\beta}'\tilde{\alpha})^{-1} - I \right) \tilde{\beta}'\tilde{c}$. The variance converges to $J \left((-\tilde{\beta}'\tilde{\alpha})^{-1} - I \right) \tilde{\beta}' \tilde{\Omega} J \left((-\tilde{\beta}'\tilde{\alpha})^{-1} - I \right) \tilde{\beta}'$. Here, X_0, X_{-1}, \dots denotes the stock price data before time 0 used to calibrate the VECM.

Proof: By Lemma 1, we transform VECM(p) into.

$$\begin{aligned}\Delta \tilde{X}_t &= \prod \tilde{X}_{t-1} + \tilde{c}_t + \tilde{u}_t \\ &= \tilde{\alpha} \tilde{\beta}' \tilde{X}_{t-1} + \tilde{c}_t + \tilde{u}_t\end{aligned}\quad (3)$$

Since $\Delta \tilde{X}_t = \tilde{X}_t - \tilde{X}_{t-1}$, Eq. (3) can be converted to

$$\tilde{X}_t = (I + \tilde{\alpha} \tilde{\beta}') \tilde{X}_{t-1} + \tilde{c}_t + \tilde{u}_t \quad (4)$$

Multiplying Eq. (4) by β' gives

$$\begin{aligned}\beta' \tilde{X}_t &= \beta' (I + \tilde{\alpha} \tilde{\beta}') \tilde{X}_{t-1} + \beta' \tilde{c}_t + \beta' \tilde{u}_t \\ &= (\beta' + \beta' \tilde{\alpha} \tilde{\beta}') \tilde{X}_{t-1} + \beta' \tilde{c}_t + \beta' \tilde{u}_t \\ &= (I + \tilde{\beta}' \tilde{\alpha}) \beta' \tilde{X}_{t-1} + \beta' \tilde{c}_t + \beta' \tilde{u}_t \\ &= (I + \tilde{\beta}' \tilde{\alpha})^t \beta' \tilde{X}_0 + \sum_{i=1}^t (I + \tilde{\beta}' \tilde{\alpha})^{t-i} \beta' (\tilde{u}_i + \tilde{c}_i)\end{aligned}\quad (5)$$

where \tilde{X}_0 is the realized value. Under the premise that $|\alpha'_\perp \Gamma \beta_\perp| \neq 0$, matrix $(I + \tilde{\beta}' \tilde{\alpha})^t$ converges to 0 as stated in Lemma 3. Therefore, Eq. (5) is equal to $\sum_{i=1}^t (I + \tilde{\beta}' \tilde{\alpha})^{t-i} \beta' (\tilde{u}_i + \tilde{c}_i)$ and is thus stationary. Now we retrieve $\beta' X_t$ by defining a $(K \times Kp)$ -dimensional matrix $J = (I_K, 0, 0, \dots, 0)$ such that

$$\begin{aligned}\beta' X_t &= J \tilde{\beta}' \tilde{X}_t = J (I + \tilde{\beta}' \tilde{\alpha})^t \beta' \tilde{X}_0 + J \sum_{i=1}^t (I + \tilde{\beta}' \tilde{\alpha})^{t-i} \beta' (\tilde{u}_i + \tilde{c}_i) \\ &\xrightarrow{t \rightarrow \infty} J \sum_{i=1}^t (I + \tilde{\beta}' \tilde{\alpha})^{t-i} \beta' (\tilde{u}_i + \tilde{c}_i)\end{aligned}\quad (6)$$

We can assume $\tilde{c}_t = \tilde{c}_{t-1} = \dots = \tilde{c}$, because we can determine the value of c_t for all t . Then, the expected value of $J \sum_{i=1}^t (I + \tilde{\beta}' \tilde{\alpha})^{t-i} \beta' (\tilde{u}_i + \tilde{c}_i)$ is equal to.

$$J \sum_{i=1}^t (I + \tilde{\beta}' \tilde{\alpha})^{t-i} \beta' \tilde{c} \xrightarrow{t \rightarrow \infty} J \left((I - (I + \tilde{\beta}' \tilde{\alpha}))^{-1} - I \right) \beta' \tilde{c} = J \left((-\tilde{\beta}' \tilde{\alpha})^{-1} - I \right) \beta' \tilde{c} \quad (7)$$

Next, since \tilde{u}_t is independently and identically distributed, we can assume.

$$\text{Var}(\tilde{u}_t) = E(\tilde{u}_t \tilde{u}_t') = E \begin{pmatrix} u'_t u_t & 0 & \dots & 0 \\ 0 & 0 & & \\ \vdots & & \ddots & \\ 0 & & & 0 \end{pmatrix} = \begin{pmatrix} \Omega & 0 & \dots & 0 \\ 0 & 0 & & \\ \vdots & & \ddots & \\ 0 & & & 0 \end{pmatrix} = \tilde{\Omega}$$

So, the variance of $J \sum_{i=1}^t (I + \tilde{\beta}' \tilde{\alpha})^{t-i} \beta' (\tilde{u}_i + \tilde{c}_i)$ is deduced to be

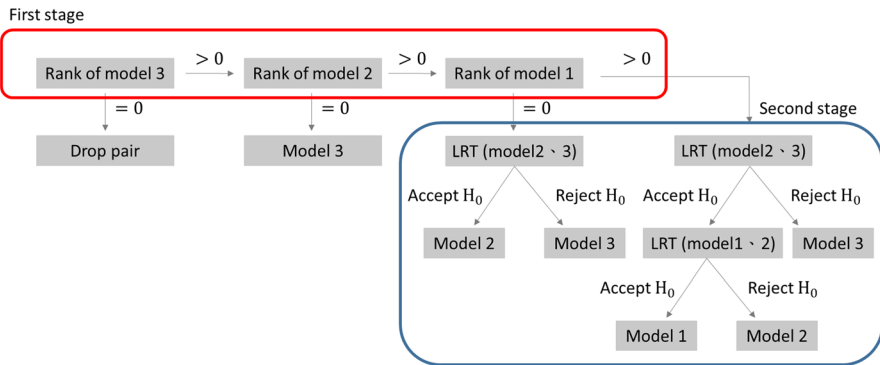


Fig. 3 Two-Step Model Selection. The first stage (in the red rectangle) uses the trace test to examine whether a VECM possesses the cointegration property. The second stage (the blue rectangle) uses the LRT to select the fittest one from the VECMs with the cointegration property determined in the first stage. (Color figure online)

$$\begin{aligned}
 \text{Var} \left(J \sum_{i=1}^t \left(I + \tilde{\beta}' \tilde{\alpha} \right)^{t-i} \tilde{\beta}' (\tilde{u}_i + \tilde{c}_i) \right) &= \text{Var} \left(J \sum_{i=1}^t \left(I + \tilde{\beta}' \tilde{\alpha} \right)^{t-i} \tilde{\beta}' \tilde{u}_i \right) \\
 &= \left(J \sum_{i=1}^t \left(I + \tilde{\beta}' \tilde{\alpha} \right)^{t-i} \tilde{\beta}' \right)' \tilde{\Omega} J \sum_{i=1}^t \left(I + \tilde{\beta}' \tilde{\alpha} \right)^{t-i} \tilde{\beta}' \\
 &\xrightarrow{t \rightarrow \infty} \left(J \left(\left(-\tilde{\beta}' \tilde{\alpha} \right)^{-1} - I \right) \tilde{\beta}' \right)' \tilde{\Omega} J \left(\left(-\tilde{\beta}' \tilde{\alpha} \right)^{-1} - I \right) \tilde{\beta}'
 \end{aligned}$$

Note that the expected value or the variance are both independent with respect to the time variable t and the initial value X_0 . This fact again confirms the stationarity.

3.3 Two-Step Model Selection

By excluding the two trend-stationary models, the remaining VECM models in Eq. (2) can be expressed in three different models (see Johansen (1995)): **Model 1**—no constant scenario,⁸ i.e., $c_t = 0$; **Model 2**—restricted constant scenario, i.e., $c_t = \alpha c_0$; **Model 3**—unrestricted constant scenario, i.e., $c_t = \mu_0$. The cointegration-based pairs trading literature either naively calibrates Model 3 or overlooks the model selection problem. However, the model selection proposed by Johansen (1995) begins the trace test⁹ from the narrowest model (Model 1) to broader models and may excessively eliminate eligible stock pairs for PTS that belong to broader models. To address this problem, we propose a two-step model selection method that

⁸ The deterministic term is zero or a constant multiple of α in the no-constant and the restricted-constant scenarios, respectively. We follow the naming of Tsay (2005) for these three models.

⁹ See Theorem 6.1 of Johansen (1995).

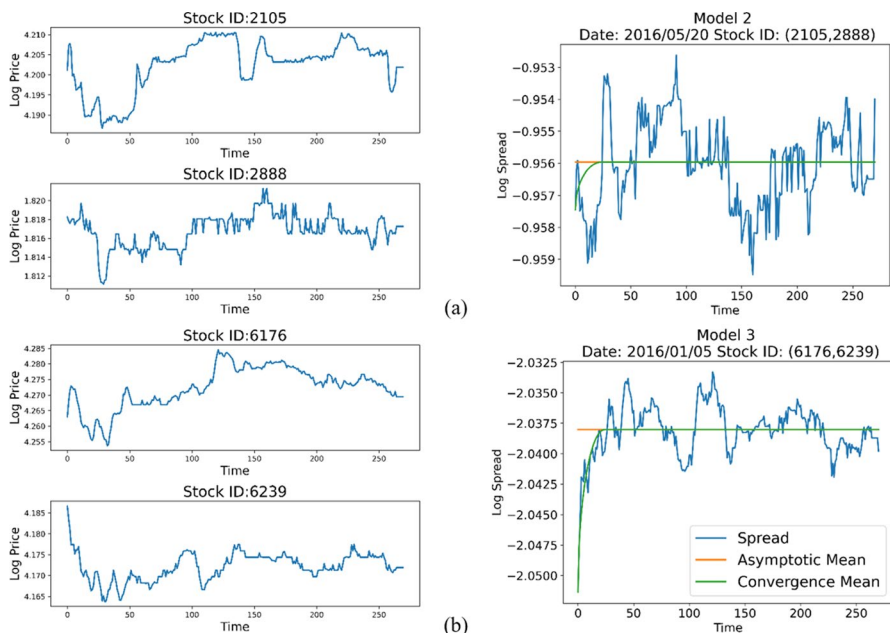


Fig. 4 Example of Model Selection and Calibration. The selected model, the ticker symbols, and the trading day are listed above each subfigure. Panel **a** calibrates Model 2 with parameters $\alpha = (-0.018, 0.052)'$, $\beta = (1, -5.993)'$, $c_t = (0.017, -0.05)'$, and $M = -0.956$. Panel **b** calibrates Model 3 with parameters $\alpha = (0.001, 0.038)'$, $\beta = (1, -2.957)'$, $c_t = (-0.003, -0.078)'$, and $M = -2.038$. (Color figure online)

uses the trace test and the likelihood ratio test¹⁰ (LRT) proposed in Johansen (1995) to determine the cointegration rank and the fittest deterministic term c_t of the VECM defined as above. The flowchart of our two-step procedure is shown in Fig. 3. In the first stage, we test for the existence of a cointegration relationship for each stock pair from the most generalized model (Model 3) down to the most restrictive one (Model 1) by estimating the rank of Π in Eq. (2) with the trace test. In the second stage, we use LRT to select the fittest VECM from the eligible models identified in the first stage. Note that the cointegration relationship exists if the rank of Π in Eq. (2) is larger than 0. If the rank obtained from calibrating Model 3 is greater than 0, a cointegration relationship exists. Thus, we exclude stock pairs whose rank is 0 because they do not possess the cointegration property under the more constrained Model 2 or Model 1. For the remaining pairs, we estimate the rank by calibrating Model 2. If the rank is 0, we consider Model 3 to be the best fitting model. Otherwise, we check the rank using Model 1. If the rank is 0, only Models 3 and 2 are suitable for describing the spread's cointegration property. In such cases, our second-stage procedure utilizes the LRT¹¹ to select the best-fitting model between Models 2 and 3.

¹⁰ See Corollary 11.2 and Theorem 11.3 of Johansen (1995).

¹¹ In Appendix A we briefly describe the use of LRT in our second stage procedure.

However, if the rank is not 0 for Model 1, all three models are suitable, and the LRT selects the best-fitting one among them.

Model selection and parameter calibration examples are given in Figs. 1 and 4. Note that the deterministic term c_t in Model 1 is 0; hence the asymptotic mean M is also 0 as in Fig. 1. The asymptotic means of both Model 2 and 3 are non-zeros. Unlike Model 2, the unrestricted deterministic term c_t of Model 3 is not a multiple of the vector α ; therefore, the stock prices and the corresponding spread patterns of these two models are distinct, as shown in panels (a) and (b) of Fig. 4.

3.4 Convergence Rate Filter

A spread with a high convergence rate to its asymptotic mean is more likely to close the position to make profits before the end of the trading day. This convergence rate can be approximated by measuring how the expected value of the spread defined in Eq. (6), that is,

$$E[\beta'x_t] = J(I + \tilde{\beta}'\tilde{\alpha})^t \tilde{\beta}'\tilde{x}_0 + \sum_{i=0}^{t-1} (I + \tilde{\beta}'\tilde{\alpha})^i \tilde{\beta}'(\tilde{c}_{t-i}) \quad (8)$$

converges to its asymptotic mean defined in Eq. (7). We denote Eq. (8) as the convergence mean since it starts from the initial value $\beta'x_0$ and converges to the asymptotic mean M with the increment of time t , as illustrated in the right panel of Fig. 1. Since a spread process with a slow convergence rate has a lower chance to close the position with positive profit, we build a convergence rate filter to filter out spreads with slow convergence rates from trading. The convergence rate is measured by how much time t_c is required for the convergence mean to approach the asymptotic mean within a tolerance error ε defined as

$$t_c := \arg \min_t |E[\beta'x_t] - M| < \varepsilon \quad (9)$$

where $E[\beta'x_t]$ is defined in Eq. (8), and the asymptotic mean M is defined in Eq. (7). Our convergence rate filter excludes spreads with a higher t_c (i.e., slow convergence rate) from trading. Our later experiments show how this filtering mechanism improves the win rate and the Sharpe ratio.

3.5 Adjustment to Approximated Simple Integer Trading Ratio

Note that we cannot buy fractions of shares since traded stocks cannot be infinitely divided, and odd lot orders are not allowed during intraday trading. However, the cointegrating vector $\beta' = [1, -\beta_1]$ determined in Eq. (2) denotes the weight of capital invested in stocks X_1 and X_2 ; thus the ratio of the number of shares for longing or shorting X_1 and X_2 at time t is a non-simple irreducible ratio vector.

$$r' = \left(\frac{1}{X_{1t}}, \frac{-\beta_1}{X_{2t}} \right)$$

where the denominators denote the prices of the two stocks at time t . In addition, we cannot amplify γ' since a large trading volume could significantly influence the stocks' prices and erode the PTS profits. To convert γ' into the closest simple integer ratio $\tilde{\gamma}'$, we find a grid point on a two-dimensional Cartesian coordinate system whose y-coordinate: x-coordinate ratio is closest to γ' by comparing the included angles using the arctangent method. To prevent large trading volumes from impacting stock prices and thereby eroding profits, we limit the maximum number of shares traded for each stock to five lots in our subsequent experiments, with one lot representing 1000 shares. $\tilde{\gamma}' \equiv (\tilde{\gamma}_1, -\tilde{\gamma}_2)$ becomes the integer number of lots for longing/shorting the stocks in the proposed PTS.

3.6 Trading Threshold Constructions and PTS Execution

On each trading day, we retrieve the price processes of all stocks from the formation period. Then we find the PTS-eligible stock pairs and obtain the corresponding ratio of the number of stock lots by the procedure in Fig. 2, after which we trade these eligible pairs during the trading period. We follow Vidyamurthy (2004) by opening the position if the spread deviates by $1.5\sqrt{\Gamma_0}$ from its the mean-reverting level M and close the position when the spread converges back to M . Specifically, we long (short) $\tilde{\gamma}_1$ lots of stock 1 and short (long) $\tilde{\gamma}_2$ lots of stock 2 at time t if $x_{1t} - \beta_1 x_{2t} \leq M - 1.5\sqrt{\Gamma_0}$ ($x_{1t} - \beta_1 x_{2t} \geq M + 1.5\sqrt{\Gamma_0}$) and close the position at time $t+i$ if $x_{1,t+i} - \beta_1 x_{2,t+i} \geq M$ ($x_{1,t+i} - \beta_1 x_{2,t+i} \leq M$). Finally, the positions may still be open at the end of the trading day. We force-close the positions to avoid the risks of cross-day portfolios. Keeping all positions closed in one trading day (i.e., intra-day trading) also decreases the Taiwan Stock Exchange's transaction cost from 0.3 to 0.15%.

We demonstrate our trading strategies in Fig. 5. The stock price processes of 1319 and 1477 during the formation period (i.e., the first 150 min) are used to examine the cointegration property and to construct a stationary spread process (denoted by the blue curve) as the trading signal for the trading period. We open the position when the spread reaches the upper threshold denoted by the green line. We observe that the sample mean tends to underestimate the mean-reverting level of the spread, resulting in positions being forcibly closed at the end of the trading day, as depicted in the upper panel. In contrast, the spread oscillates around the asymptotic mean, enabling the closure of positions for profit, as shown in the lower panel. The experiments in the following section demonstrate the superiority of using the asymptotic mean proposed in Sect. 3.2.

4 Empirical Results

To augment our contribution vis-à-vis established estimators, we first calibrate a VECM model using the historical prices that exhibit cointegration. We simulate stock price processes using this calibrated model and then employ these

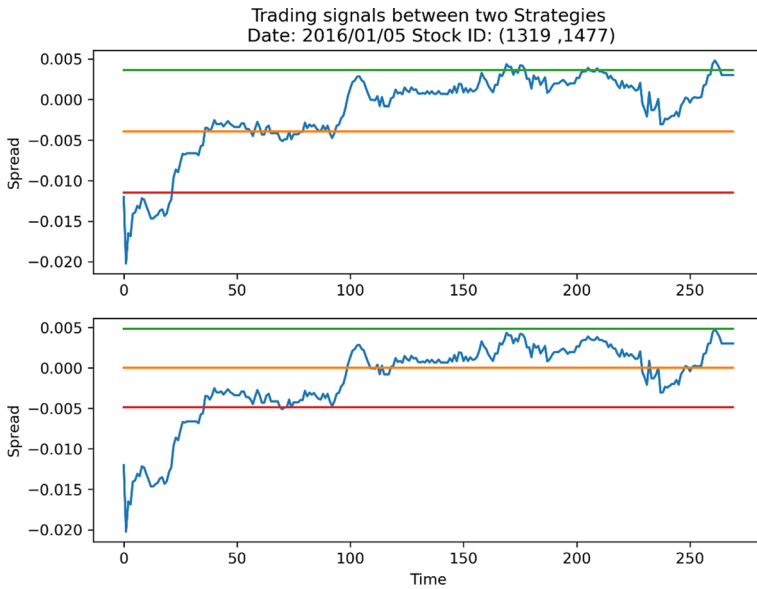


Fig. 5 Graphical Representation of PTS. The blue curve denotes the spread process by longing 3 lots of the Tong Yang Group's stock (with ticker symbol 1319) and 2 lots of Makalot Industrial Ltd. (with ticker symbol 1477) on Jan 5, 2016. The orange line denotes the mean-reverting level M estimated by the sample mean in the upper panel and the asymptotic mean in the lower panel. The green and red lines indicate the upper threshold $(M + 1.5\sqrt{\Gamma_0})$ and lower threshold $(M - 1.5\sqrt{\Gamma_0})$ for opening the positions. The blue vertical dash line divides a trading day into the first 150 min (formation period) and the remaining 120 min (trading period). (Color figure online)

simulated processes to generate a pairs trading strategy based on the cointegration property. As illustrated by the orange curves in the upper panel of Fig. 6, we utilize the stock price processes with tick numbers 6176 and 2385 from January 3, 2018. We use the data between the purple and the light blue dashed lines (i.e., the formation period) for calibration. Data to the left of the purple line are excluded from calibration to prevent excessive volatility during the market opening from impairing the calibration quality. Conversely, the data to the right of the light blue dashed line are used for implementing the PTS (i.e., the trading period). The calibration of VECM Eq. (2) yields the parameters: $\alpha = [-0.283, -0.717]$, $\beta = [0.472, 0.528]$, $c_t = [20.59, 52.22]$, $\Gamma_t = [[-0.1880.092], [0.336 - 0.237]]$, $u_t \sim N(0, [[0.0440], [0.0190.118]])$. Using these parameters, we simulate the stock price processes, represented by the blue curves. Calibrating these blue curves with Eq. (2) results in $\alpha = [-0.292, -0.708]$, $\beta = [0.483, 0.517]$, $c_t = [21.24, 51.50]$, $\Gamma_t = [[-0.2360.094], [0.477 - 0.350]]$, $u_t \sim N(0, [0.04190], [0.02220.119])$.

Now, we compare our trading simulations using models calibrated with both real and simulated prices to verify the robustness of our model. Both scenarios demonstrate similar trading ratios γ' for executing PTS, underscoring the model's consistency. In the real price scenario, the spread process crosses below the lower opening threshold in the 180th minute and returns above the mean in the 181st minute. Based

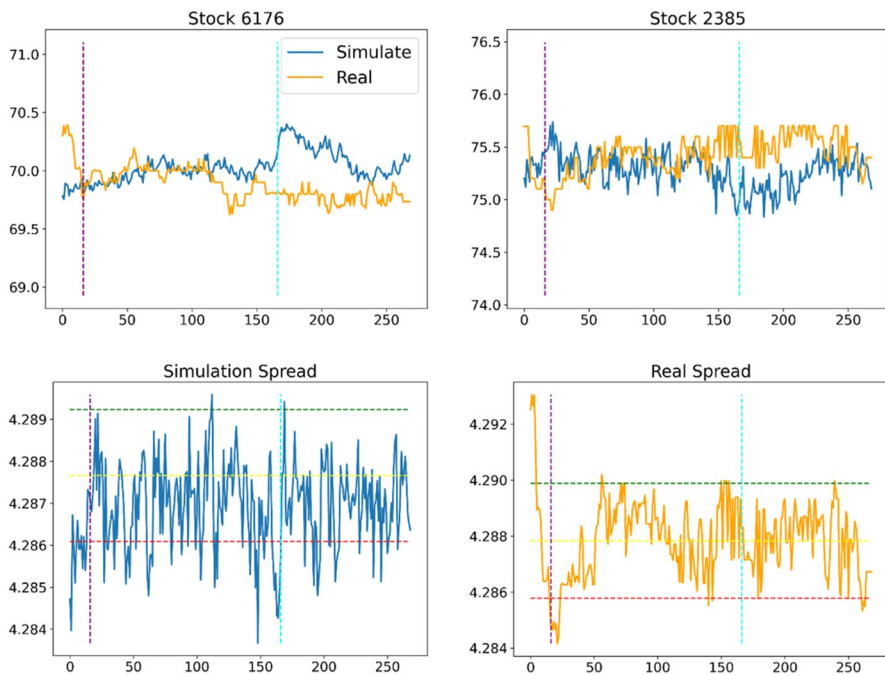


Fig. 6 The upper left and right panels display the stock prices with tick numbers 6176 and 2385, respectively. In these panels, the simulated and real prices are depicted by the blue and orange curves, respectively. The formation period is indicated between the purple and light blue vertical dashed lines, with the area to the right of the light blue line representing the trading period. The simulated and real spread processes used for PTS trading are shown in the lower left and right panels, respectively. The green, yellow, and red horizontal lines in these panels denote the upper open threshold, the mean-reverting level, and the lower open threshold, respectively. (Color figure online)

on the capital weights (0.472, 0.528) divided by the opening prices (69.7, 75.3), we obtain γ' as (0.491, 0.509). Considering the maximum profit potential and limits on the number of shares traded, we long 5 lots (i.e., 5000 shares) of the paired stock at a cost of NTD 725,000 and close the position at NTD 727,166.5, yielding a profit of NTD 2,166.5. The total transaction costs, including the Securities Transaction Tax at 0.15%, are NTD 1,090.75. Thus, the net profit for this scenario is NTD 1,075.75 (=NTD 2,166.5-1,090.75).

For the simulated price scenario, the spread crosses above the open threshold in the 170th minute and returns below the mean in the 171st minute. Using capital weights (0.483, 0.517) divided by the opening prices (70.3705, 75.3117), we obtain γ' as (0.5001, 0.4999). We short 5 lots of the paired stock in the 170th minute for NTD 728,411 and close the position in the 171st minute at NTD 726,407, resulting in a profit of NTD 2,004. The total transaction costs are NTD 1,092.617, leading to a net profit of NTD 911.383 (=NTD 2,004-1,092.617).

Our experiments illustrate the trading performance improvements by adopting the proposed asymptotic mean method in Table 1, the two-step model selection

Table 1 Trading performance using asymptotic and sample means

Trading performance										
Year	Avg PNL per trade (\$)		Win rate (%)		Normal close rate (%)		Sharpe ratio		Total trades	
	Asy	Sam	Asy	Sam	Asy	Sam	Asy	Sam	Asy	Sam
2015	0.33	0.26	89.32	85.35	87.78	82.94	0.51	0.50	3520	4982
2016	0.35	0.28	91.52	87.7	89.13	84.61	0.55	0.50	3514	4634
2017	0.53	0.39	90.63	86.34	89.07	84.00	0.70	0.60	2689	3500
2018	0.47	0.39	88.73	84.48	87.79	83.17	0.25	0.24	28,845	3510

in Table 2, the convergence rate filter method in Table 3, and the white noise test for residuals in Table 4. Finally, we examine the influence of changing the sample period duration and the number of observations on the power of the Johansen cointegration test, as discussed in Otero et al. (2022), with the results presented in Table 5. Here we first suggest quantitative indicators by which to compare the performance of different trading strategies through experiments. "Avg. PNL (profit and loss) per trade" denotes the average profit and loss for each trade in thousand New Taiwan dollars, "Win rate" denotes the percentage of trades with positive profits, "Normal close rate" denotes the ratio of spreads that converge back to the mean-reverting level before the close of the trading day, "Sharpe ratio" indicates the ratio of average trading return per trade divided by the standard deviation of the trade return, and "Total trades" denotes the number of trades over the period.

Table 1 illustrates that adopting the asymptotic mean (Asy) derived in Eq. (7) as the mean-reverting level outperforms adopting the sample mean (Sam). Specifically, the former approach has a higher average PNL per trade, win rate, normal close rate, and Sharpe ratio than the latter approach in every year of 2015–2018. However, the much larger total trades and lower normal close rates imply that the sample mean estimates a biased spread mean-reverting level and hence employs more infeasible stock pairs for PTS.

The PTS literature either ignores the cointegration model selection problem or uses the most general model, Model 3, defined in Sect. 3.3, to describe all price processes of cointegrated stock pairs. Table 2 compares the investment performance for selecting the fittest VECM using our two-step model selection described in Sect. 3.3, Johansen (1995)'s model selection, or directly calibrating Model 3 (M3). It can be observed that in each year from 2015 to 2018, our two-step model exhibits the best performance in terms of Avg. PNL per trade, win rate, normal close rate, and Sharpe ratio. By selecting appropriate cointegration models, our two-step method provides a better VECM fit, resulting in the trading of fewer infeasible pairs with higher win rates. This also explains why our model selection significantly improves investment performance.

Recall that the convergence rate of a spread process to its mean-reverting level can be approximated by how fast its convergence mean defined in Eq. (8)

Table 2 We evaluate the investment performance under different model selections

Trading performance										
Year	Avg PNL per trade (\$)		Win rate (%)		Normal close Rate (%)		Sharpe ratio		Total trades	
	MM	M3	MM	M3	MM	M3	MM	M3	MM	M3
2015	0.33/0.07	0.04	89.32/79.87	79.4	87.78/76.89	75.7	0.51/0.10	0.08	3520/2583	28,569
2016	0.35/0.05	0.14	91.52/79.72	81.5	89.13/76.44	77.9	0.55/0.08	0.33	3514/2466	25,038
2017	0.53/0.26	0.24	90.63/81.34	81.7	89.07/77.71	78.4	0.70/0.27	0.39	2689/1377	17,821
2018	0.47/-0.01	0.01	88.73/75.39	78.3	87.79/72.69	76.0	0.25/-0.01	0.01	2884/1849	20,460

In the multi-model (MM) column, we use bold and italic to denote the performance obtained by our two-step and Johansen's model selection, respectively. The mean-reverting level is set to the asymptotic mean

Table 3 Trading performance with/without using convergence rate filter under different model selection

Trading performance								
Year	M3				MM			
	No filter		Convergence filter		No filter		Convergence filter	
	Win rate (%)	Sharpe ratio	Win rate (%)	Sharpe	Win rate (%)	Sharpe Ratio	Win rate (%)	Sharpe ratio
2015	78.6	0.18	80	0.11	89/80	0.37/0.13	90/83	0.46/0.15
2016	83.9	0.33	81	0.14	92/80	0.53/0.13	92/82	0.60/0.16
2017	83.0	0.33	82	0.20	91/81	0.52/0.12	92/85	0.58/0.20
2018	77.1	0.21	79	0.15	89/75	0.32/0.03	91/79	0.37/0.11

In the multi-model (MM) column, we use bold and italic to denote the performance obtained by our two-step and Johansen's model selection, respectively. The mean-reverting level is set to the asymptotic mean

Table 4 Trading performance with/without white noise test

Trading performance										
Year	Avg PNL per trade (\$)		Win rate (%)		MDD		Sharpe ratio		Total trades	
	No test	Test	No test	Test	No test	Test	No test	Test	No test	Test
2015	0.18	0.33	83.92	89.23	1014.82	49.52	0.17	0.37	37,371	3520
2016	0.21	0.35	85.78	91.52	1575.58	45.93	0.26	0.53	36,666	3514
2017	0.43	0.53	85.94	90.63	516.05	39.98	0.26	0.52	31,221	2689
2018	0.31	0.47	83.40	88.73	1350.15	174.71	0.21	0.32	32,840	2884

The mean-reverting level is set to the asymptotic mean

approaches the asymptotic mean by time t_c defined in Eq. (9). Removing slow convergence spread processes from trading reduces the likelihood that spreads do not converge back to mean-reverting levels and are forced to close positions at the end of the trading day, potentially resulting in significant losses. A fast convergence rate implies that a divergent spread converges to the mean-reverting level more quickly and prevents structural breaks from ruining PTS benefits (see Huang et al. 2020). Indeed, our experiments in Table 3 demonstrate that trading risk decreases after removing spreads with slow convergence rates. Here "Convergence filter" denotes that we remove stock pairs whose t_c is larger than the convergence time threshold of 6 from trading; "No filter" denotes that we trade all cointegrated stock pairs without filtering. Note that filtering out slow convergence spreads significantly increases win rates and Sharpe ratios regardless of which kind of model selection is used.

In Fig. 7, we graph the impacts of varying convergence time thresholds on the average PNL per trade and the normal close rate in 2017. As a strict (i.e., lower) threshold filters out more spreads with slow convergence rates, the above two

Table 5 Analysis of impact of formation period duration and sampling frequency on power of Johansen cointegration test

Year	Comparing 1-minute sampling frequency: 150 minutes vs. 100 minutes				1-Minute vs. 2-minute sampling frequency: 150-minute duration comparison			
	Avg PNL per trade (\$)	Win rate (%)	Normal close Rate (%)	Sharpe ratio	Avg PNL per trade (\$)	Win rate (%)	Normal close Rate (%)	Sharpe ratio
2015	-0.0448	2.0673	2.4667	-0.0213	-0.2649	-11.0880	-10.2665	-0.2185
2016	-0.0372	2.5629	3.0587	-0.0187	-0.3840	-16.0982	-18.2963	-0.3031
2017	-0.1745	-0.1099	0.7114	-0.1099	-0.4917	-17.1895	-18.6880	-0.3190
2018	-0.4572	0.5943	1.0553	-0.0827	-0.8313	-16.3005	-19.0646	-0.2413

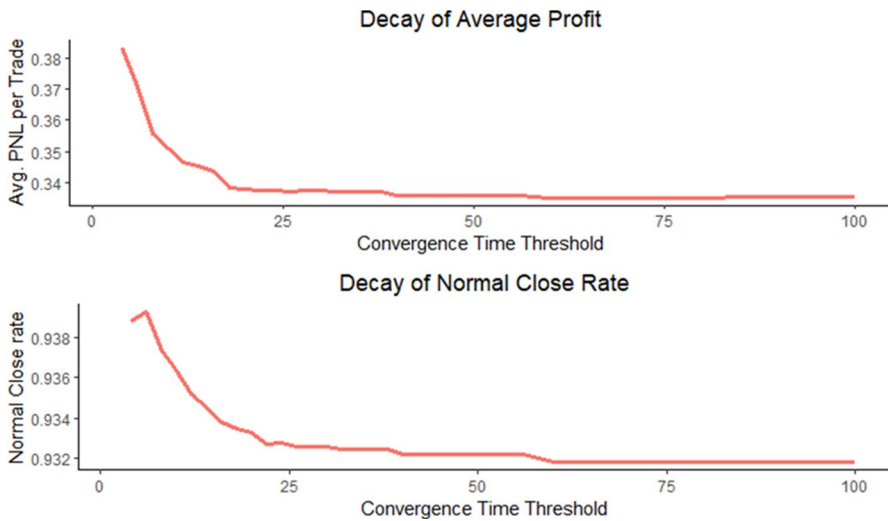


Fig. 7 Impact of Varying Convergence Time Threshold on Average PNL per Trade and Normal Close Rate

investment performance metrics also increase. In the experiments illustrated in Table 3, 6 is chosen as the convergence time threshold because these performance metrics decline significantly around 6 in Fig. 7.

The effectiveness of the VECM calibrations can be assessed by checking the quality of the residual vectors a_t by the Jarque–Bera and portmanteau tests (see Demiroglu and Kilian (2000)). If the model assumptions fit the stock prices data well, a_t should follow an independent and identically normal distribution. Otherwise, the data is inadequate for the selected VECM and might not exhibit the cointegration property. In Table 4, "Test" denotes that we remove stock pairs that fail the above two tests, and "No test" indicates that we trade all stock pairs. Although applying white noise tests significantly reduces the number of trades, it substantially increases trading quality, as seen in the increments of the average PNL per trade, the win rate, and the Sharpe ratio. It also reduces the risk, as seen in the decrements in the maximum draw down (abbreviated as MDD). Since the cointegration property is critical in PTS, we must test for white noise to ensure that the model assumptions hold.

Figure 8 displays the monthly Sharpe ratios for different PTS settings from 2015 to 2018. In the upper panel, adopting the asymptotic mean (yellow curve) as the mean-reverting level is clearly better than adopting the sampling mean (red curve). Incorporating convergence rate filters (purple curve) also improves the Sharpe ratio. Now we compare the performance of different model selections with asymptotic mean and convergence rate filters, as in the lower panel. Our two-step model selection method (purple curve) outperforms Model 3 (blue curve) and Johansen's method (green curve) in almost every month, which also confirms the robustness of our contributions.

Table 5 examines the impact of varying formation period durations and sampling frequencies on the power of the Johansen cointegration test, as reflected in trading

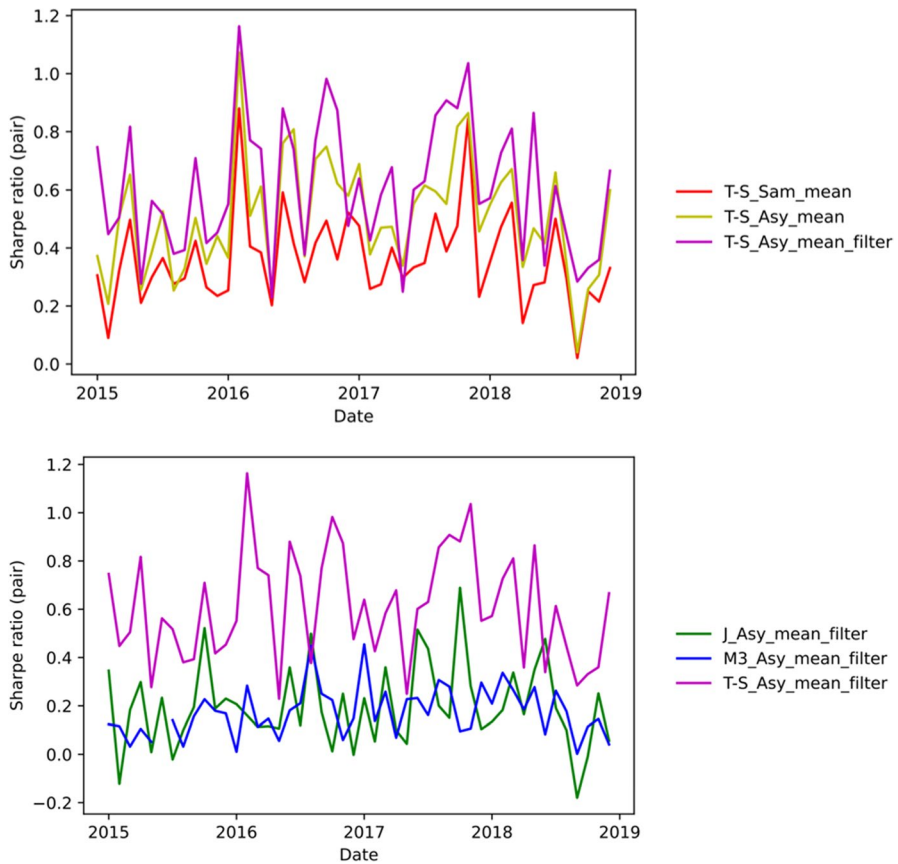


Fig. 8 4-year Monthly Sharpe Ratios. M3 denotes direct calibration of stock prices by Model 3 without model selection. J denotes Johansen's model selection. Asy and Sam_mean denote adoption of the asymptotic or sample mean as the mean-reverting level for the spread. Filter indicates adoption of the convergence rate filter. (Color figure online)

performance metrics. The benchmark adopts the above experimental settings: a 150-min formation period with a one-minute sampling frequency. Numerical results indicate the differential in trading performance between the adjusted PTS and this benchmark. The left panel shows that reducing the formation period from 150 to 100 min diminishes the test's power, as indicated by the lower Avg. PNL and Sharpe ratios. However, this reduction extends the trading period, which reduces the likelihood of forced closures at day's end and might hence increase the win rate and normal close rate. In the right panel, a decrease in sampling frequency from one minute to two minutes similarly lessens the test's power, reflected in lowered metrics across all trading performance indicators. These findings align broadly with those reported by Otero et al. (2022).

5 Conclusion

This paper analyzes the convergence properties of the spread process constructed by Johansen cointegration tests to improve PTS performance. First, we derive and use the asymptotic mean as the mean-reverting level instead of the spread's sample mean used in the literature. Our experiments show that adopting the sample mean misplaces the open and close thresholds as shown in Fig. 5, which degrades the trading performance. Second, the process during the above derivation is helpful to obtain the convergence mean used to approximate the convergence rate of a spread to its mean-reverting level. Filtering out unlikely profitable stock pairs with slow convergence rates also improves trading performance. Third, to address the joint problem of examining a stock pair's cointegration property and selecting the fittest VECM, we propose two-step model selection that makes use of the trace test and likelihood ratio test. Experiments show that this approach generates higher win rates and average profits than the Johansen cointegration approach, which is extensively used in the literature, and directly adopting the most generalized model (Model 3), which is the most general model. Removing invalid model calibrations by checking the normality of residuals also reduces trading risks. Furthermore, this paper uses the high frequency trading data of the 0050.TW and 0051.TW constituent stocks from 2015 to 2018 to examine the superiority of the proposed PTS. Unlike most studies, which execute inter-day PTS using the close price of each trading day and require sufficiently long trading (and hence backtesting) periods, our high-frequency PTS limits the portfolio holding to within one trading day to avoid overnight holding risk. Even with a limited backtesting period due to the limited availability of tick data, our high-frequency PTS still yields approximately 3000 trading opportunities per year for comparison and analysis. Also, to satisfy the constraints of no-odd-lot trading and liquidity, we convert the trading ratios of stock pairs into a nearby simple integer ratio with small magnitudes, which significantly enhances the applicability of our work for real trading.

Appendix A

We adopt the LRT proposed in Johansen (1995) to select the fittest VECM in our second-stage model selection as illustrated in Fig. 3. We test the null hypothesis Model 2 versus the alternative hypothesis Model 3 with the LRT statistic:

$$LR_{23} = T \sum_{i=r+1}^k \log \left(\frac{1 - \lambda_i^3}{1 - \lambda_i^2} \right) \quad (10)$$

where λ_i^2 and λ_i^3 are generalized eigenvalues corresponding to Model 2 and Model 3 generated by maximum likelihood estimation. The asymptotic distribution of Eq. (10) is $\chi^2(k - r)$, as illustrated in Corollary 11.2 in Johansen (1995). We choose Model 3 if the p-value is less than 0.05. If Model 1 is also an eligible candidate, we proceed to test the null hypothesis Model 1 versus the alternative Model 2 with the LRT statistic:

$$LR_{12} = T \sum_{i=1}^r \log \left(\frac{1 - \lambda_i^1}{1 - \lambda_i^2} \right) \quad (11)$$

where λ_i^1 is the generalized eigenvalue of Model 1. The asymptotic distribution of Eq. (11) is $\chi^2(r)$, as stated in Theorem 11.3 in Johansen (1995). We adopt Model 2 if the test's null hypothesis is rejected at the 0.05 significance level; otherwise we select Model 1.

Acknowledgement We would like to express our gratitude to the master and undergraduate students, Yi-Jen Luo, Chiau-Deng Mai, Hsin-Hua Chang, Heng-Sheng Kuo, Hsiu-Ming Hsu, Chung-Yu Hsu, and Bo-Kai Tasi for their assistance in this work.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by K-LW, TSD, and HHC. The first draft of the manuscript was written by Yen-Wu Ti and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by National Yang Ming Chiao Tung University. No outside funding was received for this paper.

Availability of Data and Materials Raw transaction data were generated by the Taiwan Stock Exchange. One can purchase this data from the website <https://eshop.twse.com.tw/en/>. The PTS trading data (such as stock pairs and trading ratios) generated and analyzed in this paper are available from the corresponding author on request.

Declarations

Conflict of interest All declare they have no conflict of interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References


- Avellaneda, M., & Lee, J. H. (2010). Statistical arbitrage in the US equities market. *Quantitative Finance*, 10(7), 761–782.
- Bampinas, G., & Panagiotidis, T. (2015). Are gold and silver a hedge against inflation? A two century perspective. *International Review of Financial Analysis*, 41, 267–276.
- Bampinas, G., & Panagiotidis, T. (2016). Hedging inflation with individual US stocks: A long-run portfolio analysis. *The North American Journal of Economics and Finance*, 37, 374–392.
- Başçı, E. S., & Karaca, S. S. (2013). The determinants of stock market index: VAR approach to Turkish stock market. *International Journal of Economics and Financial Issues*, 3(1), 163–171.
- Bowen, D., Hutchinson, M. C., & O'Sullivan, N. (2010). High-frequency equity pairs trading: Transaction costs, speed of execution, and patterns in returns. *The Journal of Trading*, 5(3), 31–38.

- Cummins, M., & Bucca, A. (2012). Quantitative spread trading on crude oil and refined products markets. Available at SSRN: <https://ssrn.com/abstract=1932471>
- Do, B., & Faff, R. (2010). Does simple pairs trading still work? *Financial Analysts Journal*, 66(4), 83–95.
- Do, B., & Faff, R. (2012). Are pairs trading profits robust to trading costs? *Journal of Financial Research*, 35(2), 261–287.
- Elliott, R. J., Van Der Hoek, J., & Malcolm, W. P. (2005). Pairs trading. *Quantitative Finance*, 5(3), 271–276.
- Endres, S., & Stübinger, J. (2019). A flexible regime switching model with pairs trading application to the S&P 500 high-frequency stock returns. *Quantitative Finance*, 19(10), 1727–1740.
- Engle, R. F., & Granger, C. W. (1987). Cointegration and error correction: Representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, 251–276
- Chang, H., Dai, T., Wang, K., Chu, C. and Wang, J., (2020). Improving Pair Trading Performances with Structural Change Detections and Revised Trading Strategies. In 2020 International Conference on Pervasive Artificial Intelligence (ICPAI), Taipei, Taiwan, pp. 105–109. doi: <https://doi.org/10.1109/ICPAI51961.2020.00027>
- Ferreira, L. (2008). New tools for spread trading. *Futures*, 37(12), 38–41.
- Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 19(3), 797–827.
- Hansen, P. R. (2000). The Johansen-Granger representation theorem: An explicit expression for I (1) Processes
- He, J. A., Yin, S., & Peng, F. (2023). Weak aggregating specialist algorithm for online portfolio selection. *Computational Economics*, 1–30
- Huang, S. H., Shih, W. Y., Lu, J. Y., Chang, H. H., Chu, C. H., Wang, J. Z., Huang, J. L., & Dai, T. S. (2020). Online structural break detection for pairs trading using wavelet transform and hybrid deep learning model. In 2020 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 209–216).
- Huck, N. (2009). Pairs selection and outranking: An application to the S&P 100 index. *European Journal of Operational Research*, 196(2), 819–825.
- Huck, N. (2010). Pairs trading and outranking: The multi-step-ahead forecasting case. *European Journal of Operational Research*, 207(3), 1702–1716.
- Huck, N., & Afawubo, K. (2015). Pairs trading and selection methods: Is cointegration superior? *Applied Economics*, 47, 599–613.
- Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2–3), 231–254.
- Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.
- Kilian, L., & Demiroglu, U. (2000). Residual-based tests for normality in autoregressions: Asymptotic theory and simulation evidence. *Journal of Business & Economic Statistics*, 18(1), 4–50.
- Kim, T., & Kim, H. (2019). Optimizing the pairs-trading strategy using deep reinforcement learning with trading and stop-loss boundaries. *Complexity*, 1–20, 2019.
- Krauss, C. (2017). Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 31(2), 513–545.
- Krauss, C., & Stübinger, J. (2017). Non-linear dependence modelling with bivariate copulas: Statistical arbitrage pairs trading on the S&P 100. *Applied Economics*, 49(52), 5352–5369.
- Kuo, W. L., Chang, W. C., Dai, T. S., Chen, Y. P., & Chang, H. H. (2022). Improving pairs trading strategies using two-stage deep learning methods and analyses of time (In) variant Inputs for trading performance. *IEEE Access*, 10, 97030–97046.
- Kuo, W. L., Dai, T. S., & Chang, W. C. (2021). Solving unconverged learning of pairs trading strategies with representation labeling mechanism. In CIKM Workshops
- Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1–3), 159–178.
- Li, M. L., Chui, C. M., & Li, C. Q. (2014). Is pairs trading profitable on China AH-share markets? *Applied Economics Letters*, 21(16), 1116–1121.
- Liew, R. Q., & Wu, Y. (2013). Pairs trading: A copula approach. *Journal of Derivatives & Hedge Funds*, 19(1), 12–30.
- Lin, B., & Tan, Z. (2023). Exploring arbitrage opportunities between China's carbon markets based on statistical arbitrage pairs trading strategy. *Environmental Impact Assessment Review*, 99, 107041.

- Lin, Y. X., McCrae, M., & Gulati, C. (2006). Loss protection in pairs trading through minimum profit bounds: A cointegration approach. *Journal of Applied Mathematics and Decision Sciences*. <https://doi.org/10.1155/JAMDS/2006/73803>
- Liu, B., Chang, L. B., & Geman, H. (2017). Intraday pairs trading strategies on high frequency data: The case of oil companies. *Quantitative Finance*, 17(1), 87–100.
- Liu, C., Gao, F., Zhang, M., Li, Y., & Qian, C. (2023). Reference vector-based multiobjective clustering ensemble approach for time series forecasting. *Computational Economics*, 1–30
- Lu, J.-Y., Lai, H.-C., Shih, W.-Y., Chen, Y.-F., Huang, S.-H., Chang, H.-H., Wang, J.-Z., Huang, J.-L., & Dai, T.-S. (2022). Structural break-aware pairs trading strategy using deep reinforcement learning. *Journal of Super Computing*, 78, 3843–3882.
- Maiti, D., Kumar, N., Jha, D., & Sarkar, S. (2023). Post-COVID recovery and long-run forecasting of Indian GDP with Factor-Augmented Error Correction Model (FECM). *Computational Economics*, 1–26
- Malkidis, S., & Fountas, S. (2020). Liquid fuel price adjustment in Greece: A two-stage, threshold cointegration approach. *The Journal of Economic Asymmetries*, 22, e00171.
- Otero, J., Panagiotidis, T., & Papapanagiotou, G. (2022). Multivariate cointegration and temporal aggregation: Some further simulation results. *Computational Economics*, 59, 59–70. <https://doi.org/10.1007/s10614-020-10062-w>
- 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.
- Sapna, S., & Mohan, B. R. (2023). Comparative Analysis of root finding algorithms for implied volatility estimation of ethereum options. *Computational Economics*, 1–36
- Sarmiento, S. M., & Horta, N. (2020). Enhancing a pairs trading strategy with the application of machine learning. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2020.113490>
- Stander, Y., Marais, D., & Botha, I. (2013). Trading strategies with copulas. *Journal of Economic and Financial Sciences*, 6(1), 83–107.
- Stoica, P., & Selen, Y. (2004). Model-order selection: A review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4), 36–47.
- Stübinger, J. (2019). Statistical arbitrage with optimal causal paths on high-frequency data of the S&P 500. *Quantitative Finance*, 19(6), 921–935.
- Tian, G. G., & Guo, M. (2007). Interday and intraday volatility: additional evidence from the Shanghai stock exchange. *Review of Quantitative Finance and Accounting*, 28(3), 287–306.
- Tsay, R. S. (2005). *Analysis of financial time series* (Vol. 543). Wiley
- Tsionas, M. G., Philippas, D., & Zopounidis, C. (2023). Exploring uncertainty, sensitivity and robust solutions in mathematical programming through Bayesian analysis. *Computational Economics*, 62(1), 205–227.
- Vidyamurthy, G. (2004). *Pairs trading: Quantitative methods and analysis* (Vol. 217). Wiley
- Villani, M. (2005). Bayesian reference analysis of cointegration. *Econometric Theory*, 21(2), 326–357.
- Xu, C., & Ye, Y. (2023). Optimization of asset allocation and liquidation time in investment decisions with VaR as a risk measure. *Computational Economics*, 1–27
- Yan, T., Chiu, M. C., & Wong, H. Y. (2020). Pairs-trading under path-dependent cointegration. Available at SSRN 3737985

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yen-Wu Ti¹  · Tian-Shyr Dai^{2,3}  · Kuan-Lun Wang⁴  · Hao-Han Chang²  · You-Jia Sun²

✉ Tian-Shyr Dai
cameldai@nycu.edu.tw

Yen-Wu Ti
tiyenwu@gmail.com

Kuan-Lun Wang
polyphonicared@gmail.com

Hao-Han Chang
changhaohan@nycu.edu.tw

- ¹ School of Information Engineering, Xiamen Ocean Vocational College, Xiamen 361100, China
- ² Department of Information Management and Finance, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
- ³ Risk and Insurance Research Center, National Chengchi University, Taipei, Taiwan
- ⁴ Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan