Beyond Linear Pair Trading Models: A Mixed Copula and Extended Kalman Filter Approach

Abstract

This paper proposes two novel pairs trading strategies: the Copula and Kalman Filter, utilizing Nasdaq Composite (IXIC) and Russell 2000 (RUT) Index from 2018 to 2022. Essentially, we push forward the context of traditional linearity from underlying correlation and co-integration presumption to a wider non-linear theme through extended Kalman Filter and Mixed Copula. Empirical alpha-generating evidence is obtained through both full-time analysis as well as sub-period comparison between baseline and our novel strategies. By integrating specific rules and pair selection techniques including Box-Tiao, our outcome represents that a mixed-Copula mechanism is entitled to superior performance in terms of Sharpe Ratio and annualized return, even in a bear market regime.

Keywords: Pairs Trading, Copula, Kalman Filter, Trading Strategy, Co-integration, Statistical Arbitrage

1. Introduction

Pairs Trading is well-acknowledged for its extensive application in the financial industry [1], amongst which the profitability of Distance Method proposed by Gatev et al [2] as the pioneering idea is demonstrated. This speculative strategy aims at identifying two assets (usually stocks) that are closely related in the long term historically and profiting from temporary mispricing when the two prices deviate from their equilibrium. Typically, the pair selection ensues from co-integration or correlation criteria, with trading signals triggered later by price divergence range.

Nevertheless, the rudimentary co-integration method relies heavily on the linearity assumption, and hence there exists a plethora of alternatives to amend the fact that financial data are rarely linearly associated. According to Crook and Moreira [3], excessive kurtosis is frequently observed in the equity market, leading to upper and lower tail dependence of inconsistent extent, which in turn translates to lack of co-integration sufficiency.

In this paper, we strive to develop inclusive pairs trading strategies applicable to both linear and non-linear asset relationships given market conditions of all kinds. We start with basic pair formation ideas evolving from correlation to mean-reversion. Then, we move on to constructing trading baseline yielded by co-integration and Kalman Filter based on linearity. After that, derived from Extended Kalman Filter (EKF) and Copula (both single and mixed), strategies are advanced to non-linearity, with rules and performance evaluation followed. Eventually, we compare all proposed models and discuss potential limitations for further development.

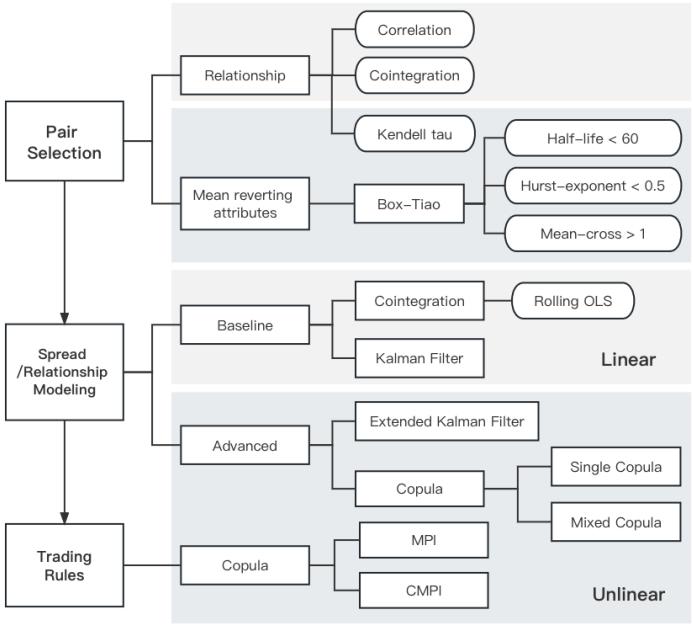


Figure 1: Flow chart of our overall methodology

2. Pairs Selection

The first step of pairs trading is to select two assets that exhibit "co-moving" attributes. In previous investigations, such attributes can be depicted by Pearson's correlation, distance approaches, co-integration tests, etc [4]. Nonetheless, given that financial assets tend to imply non-linear relationships, it is natural for us to introduce non-linear mechanisms to improve asset performance compared with

the baseline model. In this paper, two perspectives are involved: the co-moving relationship and the mean-reverting attributes of the spread.

2.1. Co-moving Relationship

Despite the sizeable application of Pearson's correlation in pairs formation, we utilize Kendall's tau which quantifies the strength and direction of association between two variables, ranging from -1 to 1. Moreover, as it is a relative-value rank-based measure rather than absolute values, it is more favorable than Pearson's correlation when two variables represent non-linear relationship. Also, since Kendall's tau is used to estimate the parameter of Copula functions, it can be directly related to our Copula strategy.

Kendall's tau can be calculated by:

$$\tau = (n^+ - n^-)/\sqrt{((n^+ + n^- + n^{\mathbf{x}})(n^+ + n^- + n^{y})}$$

where

n^+ is the number of pair tuples (x_i, y_i) and (x_j, y_j) that satisfy $(x_i > x_j \; and \; y_i > y_j)$ *or* $(x_i < x_j \; and \; y_i < y_j)$

n^+ is the pair of tuples that satisfy $(x_i < x_j \; and \; y_i > y_j)$ *or* $(x_i > x_j \; and \; y_i < y_j)$

$\mathrm{n}^{\mathbf{x}}$ is the pair of tuples in which $x_i = x_j$

$\mathrm{n}^{\mathbf{y}}$ is the pair of tuples in which $y_i = y_j$

2.2. Mean-reverting Property

A single co-moving relationship would not suffice, and hence we move on to explore the mean-reverting properties of the spread. Box-Tiao method, for its non-linear underlying hypothesis and incorporation of information about interventions or outliers, is deployed to construct an optimal hedge ratio. Then, by using the optimal hedge ratio, we calculate the historical spread and select the optimal pairs based on the following scheme:

1. Hurst exponent ≤ 0.5

This indicates the spread is an anti-persistent time series with no long-term memory.

2. Half-life ≤ 60 days

Chances are that the strategy cannot profit if it takes too long for the two assets to revert to equilibrium pricing.

3. Mean crossing ≥ 1

The number of times that the spread history crosses its mean level should be no less than one.

3. Baseline Models

Our proposal consists of two baseline models: a linear model using the co-integration method, and a Kalman Filter model. To implement the co-integration model, we utilize the Engle-Granger Two-step Method.

For the co-integration method, we assume the target pair X,Y fulfills the following relationship:

$$\log Y_t = \alpha \log X_t + \beta + v_t$$

where $\alpha > 0$ and $v_t = \log Y_t - \alpha \log X_t - \beta$ is defined as the spread of the pair. Since alpha and beta may change over time, a rolling ordinary least squares (OLS) approach and time-varying Kalman Filter (KF) are adopted to update the estimation[5].

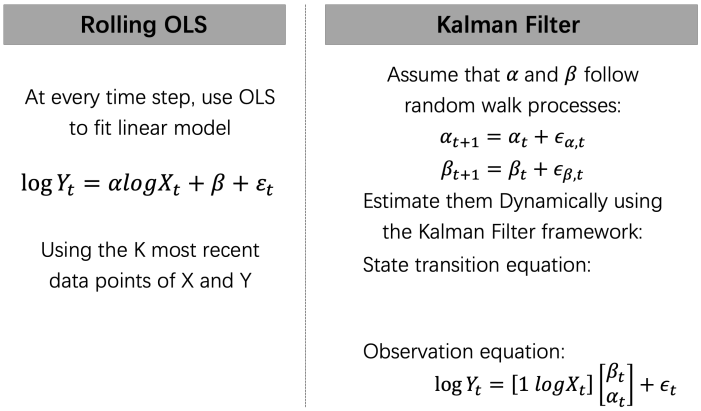


Figure 2: Comparsion between OLS and time-varying KF [6]

Step 1: Estimating the co-integration relationship

Firstly, we investigate the co-integration between two asset prices, which indicates a long-run equilibrium pricing. The parameters estimated by rolling window OLS are adopted. The residuals (ε_t) are later examined under the stationarity Augmented Dickey-Fuller (ADF) test [7]. We select 0.05 as the critical value - when the p-value is over 0.05, the unit root test is passed. Only when co-integration status is valid do we proceed to generate trading signals.

Step 2: Testing causality and forecasting

Once co-integration is established, we apply Vector Error Correction Model (VECM) to take into account the long-run relationship between the variables and short-term dynamics. The VECM is of the form:

$$\Delta y_t = \alpha_0 + \beta_0 \Delta x_t + \Gamma \varepsilon_{t-1} + \varepsilon_t$$

where Δy_t and Δx_t are the first differences of the two variables; α_0, β_0, and Γ are coefficients; ε_{t-1} is the lagged error correction term; ε_t is the error term.

In the VECM model, the adjustment coefficient Γ represents the speed at which the variables return to their long-run equilibrium after an interim shock. When the adjustment coefficient Γ is positive, the model suggests that the two assets deviate further away in response to a shock.

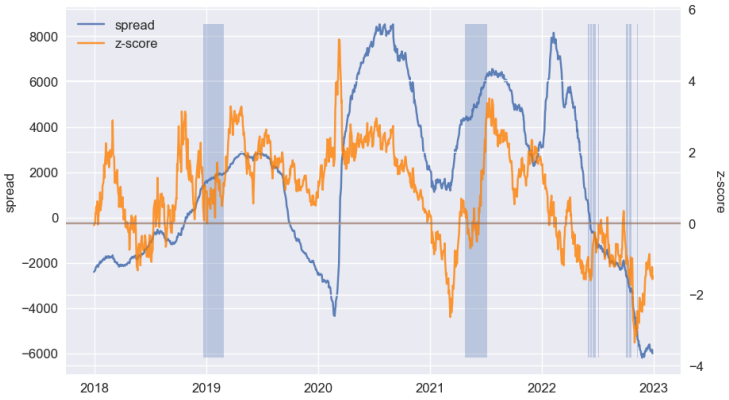


Figure 3: The spread and z-scores for IXIC and RUT

We verify the status of co-integration by checking the latest historical data sets in the rolling window. If the status is valid, we calculate the z-score of the residuals as in Figure 3, where the shaded areas are authentic co-integration periods.

Based on the mean reverting properties, we short the spread when the z-score is above the upper bound and long the spread when the z-score is below the lower bound. When the co-integration status is invalid or if the last z-score falls within the two boundaries, we do not trade. Our strategy can be visualized as in figure 4.

Note that for the co-integration, the dependent variable (y) and independent variable (x) are not interchangeable. Although from an OLS perspective, switching the two variables does not affect the significance of the coefficient, it influences the error term, which is directly related to spread modeling. Ideally, the error term shall follow a stationary and normally distributed white noise process. Therefore, we use Durbin-Watson test and Jarque-Bera (JB) test to determine which asset price should be treated as the dependent variable.

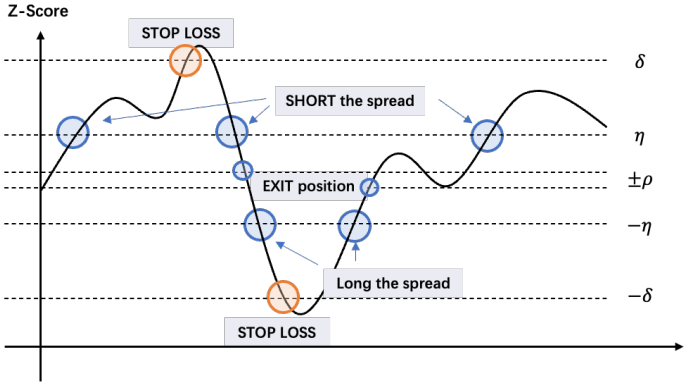


Figure 4: Illustration of signal generation

4. Nonlinear Methods

4.1. Extended Kalman Filter

Nevertheless, the Kalman Filter makes use of linear models, which is not suitable for solving nonlinear problems. Thus, we would like to introduce the Extended Kalman Filter (EKF).

In order to characterize the non-linear dependence in financial data which was previously discovered in the weekly returns of CRSP by Scheinkman and LeBaron [8], we allow $f(x)$ to be non-linear in a *Gaussian non-linear State Space model*. The EKF therefore can generate estimations to approximate the non-linear relation between two indices at each time k [9], referring to the best estimation for the prior state variable.

Instead of changing the Kalman Filter equations, the EKF solves the non-linear problem through linearization. For a given time k, we can compute the linear approximation utilizing 1st order Taylor expansion:

$$f_k(x_k) \approx f_k(a_k) + \left.\frac{\partial f_k(x_k)}{\partial x_k}\right|_{x_k = a_k} (x_k - a_{k|k})$$

where $\partial f_k / \partial x_k$ is the Jacobian matrix of function f_k assessed in $a_{k|k}$ [10].

According to Taylor expansion, we obtain the EKF equations with linearized transition model and linearized measurement model:

Despite its operational convenience and mathematical rigor, EKF has certain flaws when extensive investigations are included because it is sensitive

Algorithm 1 : Extended Kalman Filter [10]

Linearized transition model

$$x_{k+1} = f_k(x_k, u_k, w_k)$$

$$\approx f_k(\hat{x}_k, u_k, 0) + F_k(x_k - \hat{x}_k) + W_k w_k$$

Linearized measurement model

$$y_k = h_k(x_k, v_k)$$

$$\approx h_k(\hat{x}_k, 0) + H_k(\hat{x}_k - \check{x}_k) + V_k v_k$$

Step 1: Prediction

1.1. Predict the next state based on previous state

$$\check{x}_{k+1} = f_k(\hat{x}_k, u_k, 0)$$

1.2. Predict the next error convariance

$$\check{P}_{k+1} = F_k \hat{P}_k F_k^T + W_k Q_k W_k^T$$

Step 2: Correction

2.1. Compute the Kalman Gain

$$K_k = \check{P}_k H_k^T (H_k \check{P}_k H_k^T + V_k R_k V_k^T)^{-1}$$

2.2. Update the estimation through measurement

$$\hat{x}_k = \check{x}_k + K_k(y_k - h_k(\check{x}_k - 0))$$

2.3. Update the error covariance

$$\hat{P}_k = (1 - K_k H_k)\check{P}_k$$

Where

$\check{\mathbf{x}}_\mathbf{k}$ is prediction at time k

$\hat{\mathbf{x}}_\mathbf{k}$ is corrected prediction at time k

$\mathbf{F_k} = \left.\frac{\partial f_k}{\partial x_k}\right|_{\hat{x}_k, \hat{u}_k}$

$\mathbf{W_k} = \left.\frac{\partial f_k}{\partial x_k}\right|_{\hat{x}_k, \hat{u}_k}$

$\mathbf{H_k} = \left.\frac{\partial h_k}{\partial x_k}\right|_{\hat{x}_k, 0}$

$\mathbf{V_k} = \left.\frac{\partial h_k}{\partial v_k}\right|_{\hat{x}_k, 0}$

The matrices $\mathbf{F}$, $\mathbf{W}$, $\mathbf{H}$, $\mathbf{V}$ are Jacobian matrices of the system.

to linearization errors. As the approximation for a non-linear function using 1st order Taylor expansion, when non-linear function varies quickly, superior performance cannot be guaranteed. Consequently, the estimated mean may deviate from the real mean [11], while the estimated covariance may fail to capture the actual uncertainty. Furthermore, since the EKF uses the previous state to gauge the next state, the linearization errors are accumulated, resulting in estimations deviating from the actual states [9].

4.2. Copulas

4.2.1. Introduction to Copulas

Despite its advantages of capturing nonlinear relationships between assets, updating parameters dynamically, and improving strategy performance compared to the baseline models, as the underlying logic of EKF is still spread modeling, the same drawback as the baseline models exists: asset 1 and asset 2 are not interchangeable. To tackle this problem, we introduce the Copula measure to interpret the non-linear association between assets without spread modeling. This avoids estimating hedge ratios that either rely on linear relationships or are unstable over time and thus significantly enhances the model robustness.

Copula is able to separate the dependency structure from the marginal distributions in a multivariate distribution [12]. Accordingly, we can construct a multivariate distribution that portrays the dependency structure amongst asset prices or returns without explicitly assuming a specific joint distribution. In prior literature, the most frequently used Copulas are Gaussian Copula and t Copula from the Elliptical Copula family, as well as Gumbel, Clayton, and Frank Copula from the Archimedean Copula family [13]. Whilst Elliptical Copulas are centrally symmetric, some Archimedean Copulas can capture tail dependence which is helpful in seizing the dependence between financial assets. In this study, for simplicity, we only choose Copulas from the Archimedean Copula family.

To fit the Copula, we apply a Pseudo-MLE method to estimate the marginals through an Empirical Cumulative Distribution Function (ECDF) approach, and then approximate the Copula parameters via Maximum likelihood estimation (MLE). The empir-

ical CDF of the samples $y_{i,j}, ..., yn$, j is calculated by mapping each sample to:

$$\widehat{F}_{Y_j}(y) = \frac{\sum_{i=1}^{n} 1_{\{y_{i,j} \leq y\}}}{n+1}$$

and the parameter for our fitted Copula is θ_C that maximizes:

$$\sum_{i=1}^{n} \log\left[c_{\mathbf{Y}}\left(\widehat{F}_{Y_1}(y_{i,1}), \ldots, \widehat{F}_{Y_d}(y_{i,d}) \mid \boldsymbol{\theta}_C\right)\right]$$

Applying the ECDF approach will save estimating parameters for marginal CDFs and simplify the optimization process.

4.2.2. Single Copulas

We start our Copula mechanism with a single Copula strategy. First, we transfer both asset prices in the training data to quantiles using ECDF and fit them into single Copulas. Then, with the partial derivative calculated from the Copula model, we obtain the conditional probabilities $P(U_1 < u_1 \mid U_2 = u_2)$ and $P(U_2 < u_2 \mid U_1 = u_1)$, which we refer to as the mispricing index (MPI). When the former is relatively small, it implies that asset 1 is undervalued, while when the latter is relatively large, it infers asset 2 is overvalued. In this case, the two conditional probabilities represent the same directions for spread trading. Therefore, our signals for the single Copulas are generated by:

If $P(U_1 < u_1 | U_2 = u_2) \leq \delta_-$
and $P(U_2 < u_2 | U_1 = u_1) \geq \delta_+$, long the spread;
If $P(U_1 < u_1 | U_2 = u_2) \geq \delta_+$
and $P(U_2 < u_2 | U_1 = u_1) \leq \delta_-$, short the spread;

When both conditional probabilities cross the boundary at 0.5, the model indicates there no longer exists mispricing and we exit.

However, according to Silva [13] and Rad [7], this method is not robust. Instead, inspired by Xie et al [14], we introduce the CMPI approach.

Improved Signals

The Cumulative Mispricing Index (CMPI) is an attempt to improve the performance of the single Copula strategy. While in the MPI method we used signals from mispricing in a single day, the CMPI is capable of accumulating mispricing information during a period. Specifically, based on the MPI for two assets,

$$MPI_t^{X|Y} = P\left(R_t^X < r_t^X \mid R_t^Y = r_t^Y\right)$$

$$MPI_t^{Y|X} = P\left(R_t^Y < r_t^Y \mid R_t^X = r_t^X\right)$$

We construct two flags:

$$FlagX^*(t) = \sum_{s=0}^{t}\left(MPI_s^{X|Y} - 0.5\right)$$

$$FlagY^*(t) = \sum_{s=0}^{t}\left(MPI_s^{Y|X} - 0.5\right)$$

4.2.3. Single Copulas Strategy

Using the CMPI method, we take a long position in asset X and a short position in asset Y at time t if $FlagX^*(t)$ exceeds an upper boundary and $FlagY^*(t)$ falls below a lower boundary, and vice versa. Once the position is open, we hold and wait until the CMPI converges to 0. However, if the Flags become over high or low, we consider the current distribution to be void and is upon reevaluation. Under this circumstance, we close the position and set the CMPI to 0.

We select Clayton, Frank, and Gumbel Copulas separately to implement the strategy. The outcomes are as follows:

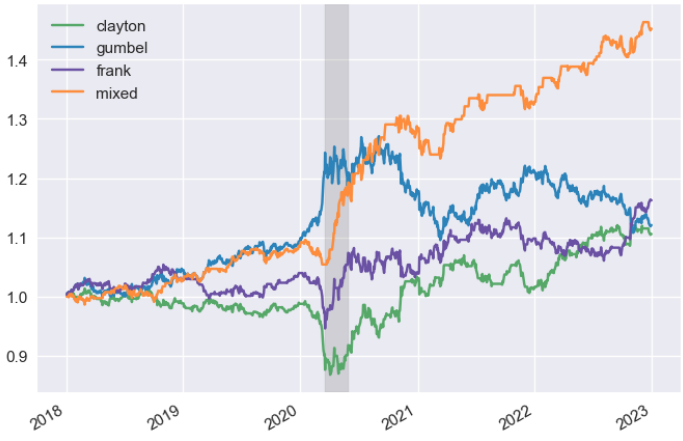


Figure 5: Copula Backtesting Results

We can see that none of these strategies can profit steadily and are all subject to drawdowns.

In order to examine the root causes for the above problems, we turn to the Clayton Copula, which is

a commonly used method in the financial sector due to its effectiveness in accurately capturing lower tail dependence. Our findings show that the strategy experienced significant losses in the months followed by the circuit breaker on March 18, 2020, whose potential cause is its disparity between the estimated Copula distribution and the actual yield distribution.

In Figure 6, the left plot showcases the returns distribution of two assets during 252 trading days before March 18, 2020, and additionally illustrates the Clayton Copula. It is intuitive that the actual data has produced upper dependencies, which Clayton fails to capture effectively.

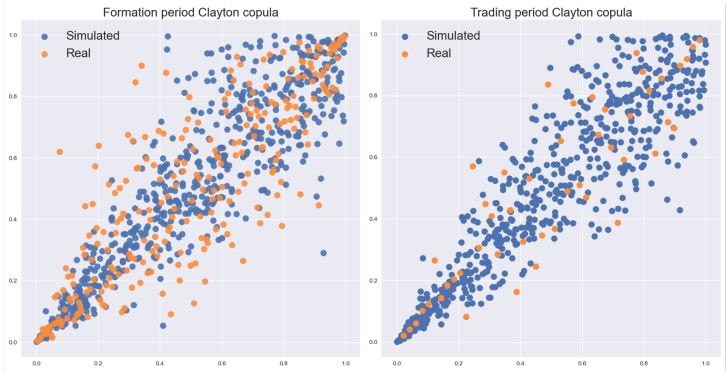


Figure 6: Clayton Simulated Results

The right plot displays the returns distribution of two assets over 60 trading days after March 18, 2020. It indicates that market panic following the circuit breaker led to a rapid increase in terms of dependency between two tails, resulting in a shuttle shape with narrow ends. However, since Clayton does not allow for upper tail dependency, this deviates from the actual distribution and may result in signal bias, eventually causing strategy losses.

When selecting the appropriate Copulas, a single Copula may not be sufficient as the market does not consistently cluster at a specific tail, and individual Copulas are either asymmetric or can only depict a specific tail. Furthermore, Copulas inevitably deviate over time, even if their parameters are updated. This is one of the reasons why each of the three strategies experiences drawdowns at different times.

To address this problem, we propose the use of a mixed Copula.

4.2.4. Mixed Copula

Mixed Copula is the combination of two or more Copulas through a weighted sum. A mixture of different Copulas allows the model to grasp more sophisticated asymmetrical tail dependence. A mixed Copula has two sets of parameters: parameters for each component Copula θ_K, and weights for each Copula ω_i. In our study, we present the Clayton-Frank-Gumbel mixed Copula, which is expressed as:

$$C_{mix}(u_1, u_2; \theta, \omega) := \omega_C C_C(u_1, u_2; \theta_C) + \omega_F C_F(u_1, u_2; \theta_F) + \omega_G C_G(u_1, u_2; \theta_G)$$

Where θ_C, θ_F, and θ_G represent parameters for individual Copulas, ω_C, ω_F, and ω_G represent their positive contribution and have a restricted sum to one.

We compute a Canonical Maximum Likelihood Estimator (CMLE) using pseudo-observations [15], and estimate the optimized parameters on a rolling basis with the SLSQP optimizer in the SciPy library. Our likelihood function is:

$$l(\theta, \omega) = \sum_{t=1}^{T} log[\sum_{k=1}^{3} \omega_k c_k(u_{1,t}, u_{2,t}; \theta_k)]$$

with the given constraint $\sum_{i=1}^{3}(\omega_i) = 1$.

We fit the mixed Copula to the same dataset as in the previous context, and arrive at the following parameters: $\omega_C = 0.38$, $\omega_F = 0.00$, $\omega_G = 0.62$, $\theta_C = 3.60$, $\theta_F = 5.87$, and $\theta_G = 3.12$.

The visualization result is shown in figure 7:

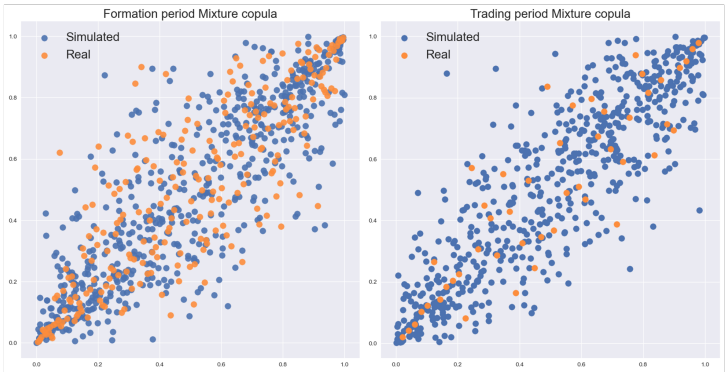
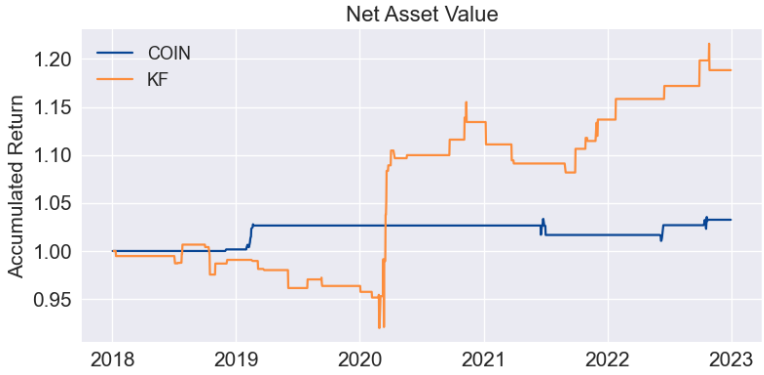


Figure 7: Mixed Copula Simulated Results

The mixed Copula successfully captures the aggregation of both tails. This is achieved by utilizing the Clayton Copula to represent lower tail dependency



Baseline Portfolio Performance

	Cointegration	Kalman Filter
Sharpe Ratio	0.52	0.52
Annualized Return	0.01	0.04
Annualized Volatility	0.01	0.07
Max Drawdown	0.02	0.09
Win Rate	0.62	0.54
Average Hoding Period	4.62	1.15

Figure 8: Baseline Results

and the Gumbel Copula to represent upper tail dependency, achieving an overall better reflection of the market's current regime.

4.2.5. Mixed Copula Strategy

Trading Signals Generation

Similar to the single Copula methodology, we apply CMPI to generate trading signals. Additionally, we include the "VolChange" indicator, which displays extreme market fluctuations. If the volatility changes by more than 10% between two consecutive days, we consider this a remarkable shift in the market environment. In such a scenario, the previously fitted Copula may not be able to adjust in time to the new market conditions, and we may need to close our existing positions.

To avoid confusion, we use the abbreviations "OB" to stand for the boundary for opening a position, and "SLB" for the stop loss boundary.

In summary, we will long X short Y at time t if all following conditions hold true:

- SLB > $FlagX^*(t)$ > OB
- −OB> $FlagY^*(t)$ > −SLB
- $VolChange < 10\%$

and vice versa.

We will close the position as long as one of the following conditions is true:

- One of the Flags has returned to zero
- One of the Flags exceeded the SLB range
- $VolChange > 10\%$

When we close a position, we set the CMPI of the pair back to 0.

According to Xie [14], we set the Stop Loss Boundary to 2 and look for the optimal Open Boundary through grid search.

Model Re-estimation

We follow the work by Gatev et al [2] to conduct a five-year backtest, including both formation and trading periods. A formation period is defined as a 12-month period, i.e. 252 days. Historical data from the formation period are used for estimation of the parameters. A trading period is defined as a 3-month period, i.e. 60 days.

At the initiation of each trading period, we re-estimate the copula parameters using the data from the previous period, so as to account for any changes in the relationship between the two assets.

We also keep an eye on the "VolChange" indicator. If $VolChange$ rises above 10%, we not only close our positions, but also update the model. Additionally, to increase the model's responsiveness to recent and significant changes, we shorten the formation window to 126 trading days during these scenarios.

5. Backtest Result

5.1. Baseline Result

The co-integration method has limitations for the baseline strategy, among which a notable one is that co-integration may not always be valid, resulting in only 8 trades over the five-year backtesting period. Additionally, signals can only be generated when

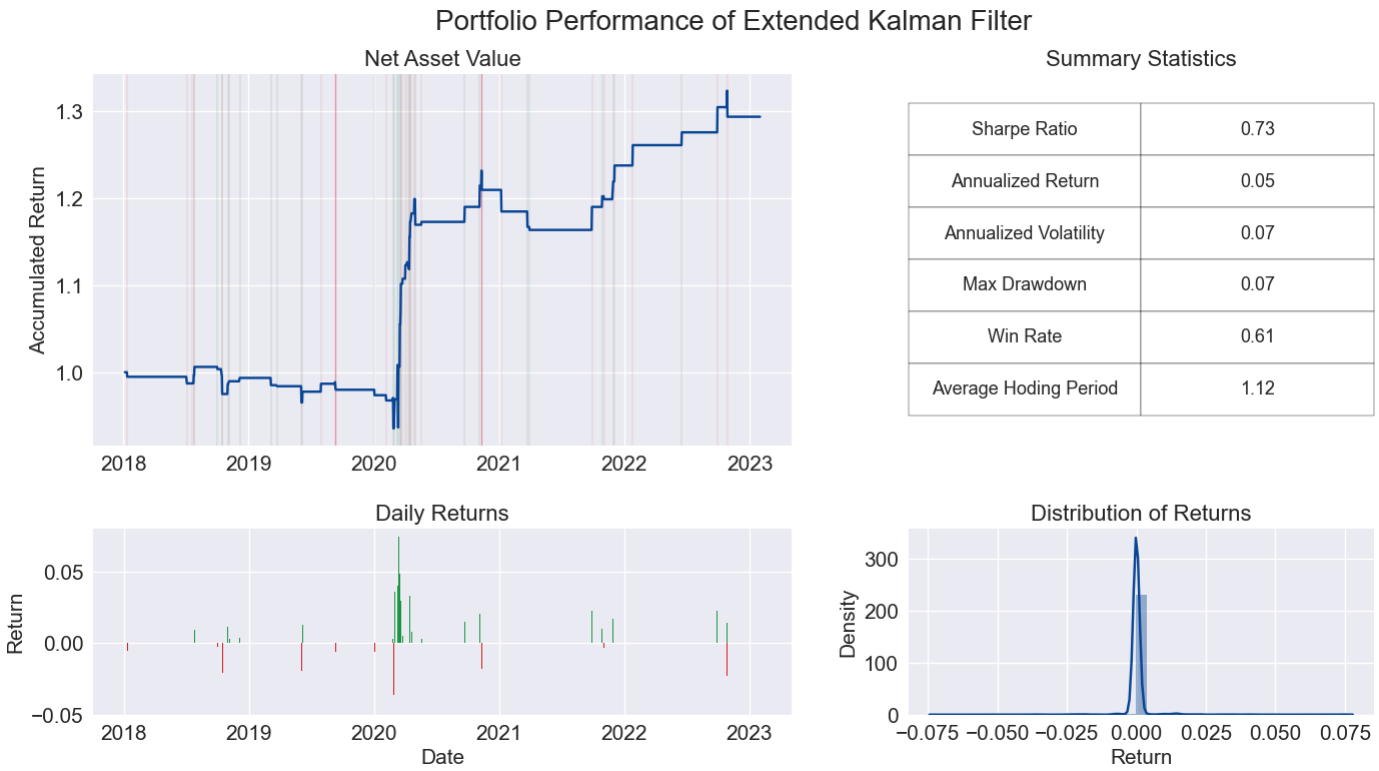


Figure 9: Extended Kalman Filter Backtesting Results

there is a deviation in asset prices from their long-term equilibrium, and the baseline strategy can only profit from the spread series crossing between the open-position threshold and the exit-position threshold, conditional on the stop-loss not triggered. As a result, the profit potential of the baseline strategy is to a large extent limited, especially when dynamic pair selection is not possible.

5.2. *Extended Kalman Filter Result*

The EKF model delivers a Sharpe ratio of 0.73 over a five-year backtesting period, higher than the KF one. Additionally, the EKF model generates a higher average return of 5.3% and a lower maximum drawdown of 0.07. The EKF model's outperformance over KF is likely due to its ability to employ higher-order approximation to estimate nonlinear functions. Furthermore, we implement grid search to optimize profits: we set the threshold to 2 and rolling window size to 110. Overall, the resulting matrices suggest that the EKF model combined with grid search is an effective tool to reduce estimation errors and thus generating profitable signals.

5.3. *Mixed Copula Result*

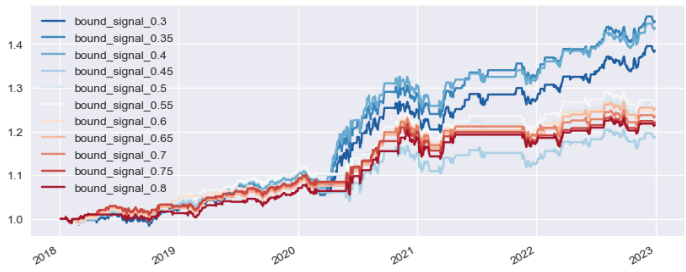


Figure 11: Mixed Copula Backtesting Grid Search Results

Over the five-year backtesting period, our mixed Copula strategy achieves a Sharpe ratio of 1.4. Figure 10 reveals that this strategy involves higher trading frequencies than the baseline with an average holding period of 9.42 days. The mixed Copula model outperforms all three single Copulas in terms of both profitability and stability.

One reason for mixed Copula's superior performance is that it balances the strengths of the three Copulas and adapts to various market conditions. Moreover, the mixed Copula strategy automatically adjusts the weights of components according to market changes. For instance, as Figure 12 shows, right

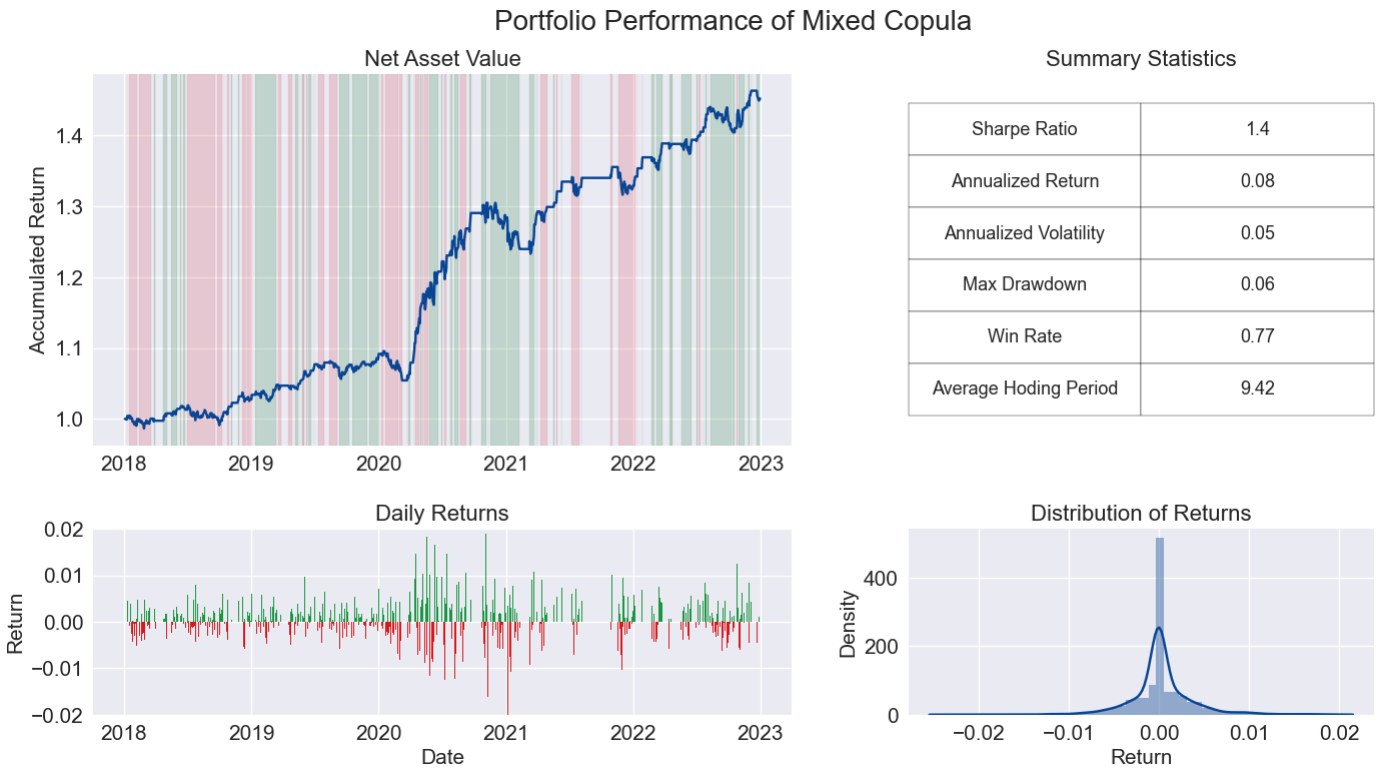


Figure 10: Mixed Copula Backtesting Results

before the bear market in 2020, the weights of Clayton were increased, followed by its gradually decreasing in favor of Frank Copula later when the market changed, as Frank had been more potent in the past year.

The mixed Copula strategy also experiences lower volatility due to the balancing effect of the three Copulas since the net asset value of strategies utilizing Clayton Copula and Gumbel Copula tend to move in opposite directions.

6. Conclusions and discussions

In this article, we develop an integrated Copula framework and an extended Kalman Filter which purports to profit in pairs trading regardless of market regimes. Commencing from pair formation, we suggest that correlation and co-integration are subject to linearity premise, which may obfuscate the right pair and spread, and therefore we put forward with mean-reverting property utilizing Box-Tiao.

To dynamically analyze the non-linear state space model with linear restriction on the observation matrix, in EKF, we internally estimate the observation covariance and input of the Jacobian transition matrix to optimize and restrict the updating process. In this way, EKF provides a closer estimation of spread (Sharp Ratio of 0.73 and return of 5% over a four-year period) by allowing non-linear dependency.

Using Copula models can be helpful in describing the non-linear relationship between two variables for pairs trading. However, relying on a single copula strategy may not be robust due to its fixed tail dependency structure, which is one of the reasons for large drawdowns. To address this challenge, we introduce a mixed Copula approach that combines three Copulas through a weighted sum.

For this study, we employ the CMLE with pseudo-observations to estimate the parameters of the mixed copulas and then move forward to the CMPI to generate trading signals. Additionally, we monitor volatility to re-estimate the parameters and adjust the window size for re-evaluation. Eventually, mixed Copula achieves a Sharpe ratio of 1.4 and an annualized return of 8%.

Through multi-dimensional evaluation, our empirical results showcase that both non-linear approaches are superior to the conventional linear baselines in terms of Sharpe Ratio, volatility, and maximum drawdowns, while the mixed Copula (constituted of: Clayton, Frank, and Gumbel) excels. Furthermore, we are able to retain sustainable profitability, and

Figure 12: Mixed Copula Weights and Thetas

even survive unexpected unfavorable market conditions.

7. Further Studies

There are certain alternatives to refine our methodology though. For example, our mixed Copula model involves filtering trading periods based on historical volatility in a rolling window. However, this approach could be improved by incorporating exponentially weighted moving average (EWMA) volatility or utilizing time series volatility forecasting, such as the GARCH model proposed by Bollerslev in 1986 [16].

Also, the CMPI method in our paper includes a hyperparameter that imposes arbitrary limits. Nonetheless, it is notable that CMPI is highly model-dependent. In our study, the hyperparameter is fixed throughout the entire backtesting period, although the optimal value and signal generation sensitivity should be contingent on different market environments. Complex methods like reinforcement learning can be carried out to fulfill this goal with a larger dataset, preferably intraday stock data.

A further extension to our work is to consolidate machine learning to determine market trigger points. Events resulting in two assets departing from their equilibrium can both be temporary or permanent: the former is represented by the U.S. equity market circuit breakers in 2020 and FTX collapse in 2022, while China's calling off pegging CNY to USD in 2005 serves as a sample for the latter.

Event-driven statistical abnormality can be analyzed and manifested using machine learning models and natural language processing (NLP). For instance, Petropoulos and Siakoulis [17] adopted an adaptive NLP sentiment index together with XGBoost mechanism to predict real time equity market fluctuations provoked by Federal Reserve monetary speeches. Likewise, when market-transforming events take place, trading signals can be generated to remedy appropriate positions movement and market timing in our strategy. Additionally, event-driven methods can inform investors of significant large-scale abnormalities (e.g., COVID-19) in a timely manner, and in turn, adjust portfolio construction case-by-case to avoid extreme model tail risks.

References

[1] G Zhang. *Pairs Trading with Nonlinear and Non-Gaussian State Space Models. arXiv: Portfolio Management*. 2020.

[2] Evan Gatev, William N Goetzmann, and K Geert Rouwenhorst. "Pairs trading: Performance of a relative-value arbitrage rule". en. In: *Rev. Financ. Stud.* 19.3 (2006), pp. 797–827.

[3] Jonathan Crook and Fernando Moreira. “Checking for asymmetric default dependence in a credit card portfolio: A copula approach”. en. In: *J. Empir. Finance* 18.4 (Sept. 2011), pp. 728–742.

[4] Christopher Krauss and Johannes Stübinger. “Non-linear dependence modelling with bivariate copulas: statistical arbitrage pairs trading on the S&P 100”. en. In: *Appl. Econ.* 49.52 (Nov. 2017), pp. 5352–5369.

[5] A Tourin and R Yan. *Dynamic Pairs Trading Using the Stochastic Control Approach. Advanced Risk & Portfolio Management® Research Paper Series*. 2012.

[6] Wong Wen Yan, Ko Chung Wa, and Tham Guang Yao. “Pairs Trading with Machine Learning”. In: (Apr. 2019).

[7] Hossein Rad, Rand Kwong Yew Low, and Robert W Faff. “The profitability of pairs trading strategies: Distance, cointegration, and copula methods”. en. In: *SSRN Electron. J.* (2015).

[8] José A Scheinkman and Blake LeBARON. “Nonlinear dynamics and stock returns”. In: *Cycles and Chaos in Economic Equilibrium*. Princeton University Press, Feb. 2021, pp. 446–474.

[9] E A Wan and R Van Der Merwe. “The unscented Kalman filter for nonlinear estimation”. In: *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*. Lake Louise, Alta., Canada: IEEE, 2002.

[10] Simon Haykin. *Kalman Filtering and Neural Networks*. en. Ed. by Simon Haykin. Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control. Nashville, TN: John Wiley & Sons, Sept. 2001.

[11] Chun-Hao Chen, Wei-Hsun Lai, Shih-Ting Hung, et al. “An advanced optimization approach for long-short pairs trading strategy based on correlation coefficients and Bollinger Bands”. en. In: *Appl. Sci. (Basel)* 12.3 (Jan. 2022), p. 1052.

[12] Rong Qi Liew and Yuan Wu. “Pairs trading: A copula approach”. en. In: *J. Deriv. Hedge Funds* 19.1 (Feb. 2013), pp. 12–30.

[13] Fernando A B Sabino da Silva, Flavio A Ziegelmann, and João F Caldeira. “A pairs trading strategy based on mixed copulas”. en. In: *Q. Rev. Econ. Finance* 87 (Feb. 2023), pp. 16–34.

[14] Wenjun Xie, Rong Qi Liew, Yuan Wu, et al. “Pairs trading with copulas”. en. In: *J. Trading* jot.2016.2016.1.048 (June 2016).

[15] Zongwu Cai and Xian Wang. “Selection of mixed copula model via penalized likelihood”. en. In: *J. Am. Stat. Assoc.* 109.506 (Apr. 2014), pp. 788–801.

[16] Richard T Baillie, Tim Bollerslev, and Hans Ole Mikkelsen. “Fractionally integrated generalized autoregressive conditional heteroskedasticity”. en. In: *J. Econom.* 74.1 (Sept. 1996), pp. 3–30.

[17] Anastasios Petropoulos and Vasilis Siakoulis. “Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique”. en. In: *Cent. Bank Rev.* 21.4 (Dec. 2021), pp. 141–153.