28 February 2023

# Copulas and Dynamic Hedging Techniques in Pairs Trading

*0. Abstract*

Pairs trading is a relative value investment strategy that seeks to identify 2 funds or indices with similar characteristics whose equity securities are currently trading at a price relationship that is out of their historical trading range. This investment strategy will entail buying the undervalued security while short selling the overvalued security based on mean reversion. In this project, we utilize Kendall Tau coefficient and sum of squared difference to examine the correlations between each pair and choose the highest-scored pairs. Next, we adopt Copula to get a joint quantile density function and generate entrance and exit signals. Lastly, we back-test and update parameter estimation.

Implementations

*1. Data Collection and Data cleaning*

To implement the project, we first acquire the historical daily stock prices of thirteen equity indices (DIA, FEC, IJH, IWD, IWF, IWM, IWN, IWO, MDY, OEF, QQQ, SPY, and VTI). We collect data from Yahoo Finance and mainly adopts “Adj.Close” price. To simplify our analysis, we further combined the “Adjusted Close Price” of all indices into a universe data frame. We also calculated the daily returns of all indices and combined them into another universe data frame.

The major step in our data-cleaning process is dealing with null values. Since equity indices are listed on different dates, their histories are different. The date of our universe data frame follows the equity with the longest history; thus, null values can emerge in the data frame. To avoid it, we filter our data by only selecting dates after 2014/12/31.

In our later analysis, we will do a part called volatility proxy, which requires daily “High” and daily “Low” in the calculation. Therefore, we iterate the above steps and obtained a data frame with daily “Highs” and daily “Lows” for all indices. Based on those two prices, we also calculated the daily range (ln (High/Low)) column for each index. All daily range data is stored in another data frame.

So far, we have four data frames, one contains all daily adjusted close prices, one contains all daily returns, one contains all daily highs and daily lows, and one contains all daily ranges.

*2. Get Correlations of the Equity Indices*

*2.1. The Sum of Squared Differences (SSD)*

Gatev et al. (2006)’s work can be considered as the baseline for distance-based selection criteria. It proposes to choose the pair that minimizes the distance criterion. In distance method, we

calculate the sum of squared differences between the returns of different combinations of index pairs during the formation period.

However, this approach may not be consistent with the goal of finding potentially profitable pairs, as a zero spread pair would be considered optimal under this criterion. A good pair for trading should have high spread variance and strong mean-reversion properties, which would provide trading opportunities. Since the SSD criterion does not consider these requirements, it may result in the formation of pairs with low spread variance and limited profit potential.

*2.2. Kendall Tau*

In correlation methods, Kendall tau is a statistical measure used to assess the degree of association or similarity between two variables, such as the returns of two stocks that are being traded as a pair. It can help identify potentially profitable pairs based on their historical relationship. Given the pairs $(X_i, Y_i)$ and $(X_j, Y_j)$, then: $\frac{Y_j - Y_i}{X_j - X_i} > 0$, pair is concordant; $\frac{Y_j - Y_i}{X_j - X_i} < 0$, pair is discordant; $\frac{Y_j - Y_i}{X_j - X_i} = 0$, pair is considered a tie. The formula to calculate Kendall's Tau, often abbreviated τ, is as follows, with $N_c$ and $N_d$ denoting the number of concordant pairs and the number of discordant pairs:

$$\tau = \frac{N_c - N_d}{N_c + N_d}$$

*3. Get Dynamic Hedge Ratios*

Hedge ratio describes the amount of instrument B to purchase or sell for every unit of instrument A. The hedge ratio can refer to a dollar value of instrument B, or the number of units of instrument B, depending on the approach taken. To calculate the hedge ratio between two indices, the conventional way is performing OLS regression, i.e., Y = $\beta$X+$\alpha$, where Y and X are daily prices of the two indices. However, we want to explore the time-varied estimates of the hedge ratio, since we assume there are certain market events that disrupt the mean-reversion trend of the pairs. Hence, we adopt OLS, Rolling OLS (simple OLS with fixed window size), as well as Kalman Filter and Volatility proxy approaches.

*3.1 Kalman Filter*

The Kalman Filter utilizes the state space mode, it essentially treats the "true" hedge ratio as an unobserved hidden variable and attempts to estimate it with "noisy" observations. Filtering means estimates the current value of the state from past and current observations. Define the state equation: $\theta_t = G_t\theta_{t-1} + \omega_t$, and the observation $y_t$ is a linear combination($F_t$) of the current state ($\theta_t$)and noise($v_t$): $y_t = Ft^T \ \theta_t \ + v_t$, where $\theta_t \sim N(m_0, C_0)$; $v_t \sim N(0, V_t)$, and$\theta_t \sim N(0, W_0)$. Remind Bayer's Rule is given by: P(H|D) = P(D|H)*P(H)/ P(D), then we can apply it to this situation: $P(\theta_t|D_{t-1,yt}) = \frac{P(y_t|\theta_t)*P(\theta_t|D_{t-1})}{P(y_t)}$.

What does this equation mean? It says that the updated probability of obtaining a state $\theta_t$ given our observation $y_t$ and previous data $D_{t-1}$, is equal to the likelihood of seeing an observation

$y_t$ given the current state $\theta_t$multiplied by the previous belief of the current state or prior, given only the precious data $D_{t-1}$, normalized by the probability of seeing the observation $y_t$.

We specify these distributions: $\theta_t|D_{t-1}$ ~ N($a_t$,$R_t$); $y_t|\theta_t \sim N(Ft^T\theta_t$, Vt), and $\theta_t|D_t \sim N(m_t,C_t)$.

Now we can take expected value of the observation tomorrow, given our knowledge of the today:

$$\begin{aligned} E[y_{t+1}|D_t] &= E[F_{t+1}^T\theta_t + v_{t+1}|D_t] \\ &= F_{t+1}^T E[\theta_{t+1}|D_t] \\ &= F_{t+1}^T a_{t+1} \\ &= f_{t+1} \end{aligned}$$

Similarly, we can derive the variance:

$$\begin{aligned} \mathbf{Var}[y_{t+1}|D_t] &= \mathbf{Var}[F_{t+1}^T\theta_t + v_{t+1}|D_t] \\ &= F_{t+1}^T \mathbf{Var}[\theta_{t+1}|D_t]F_{t+1} + V_{t+1} \\ &= F_{t+1}^T R_{t+1}F_{t+1} + V_{t+1} \\ &= Q_{t+1} \end{aligned}$$

Now that we have the expectation and variance of tomorrow's observation, given today's data, we are able to provide the general forecast for k steps ahead: $y_{t+k}|D_t \sim N(f_{t+k|t},Q_{t+k|t})$. The implementation of Kalman Filter in Python is relatively easy, we simply use the KalmanFilter class from pykalman library. We set the initial state mean to be zero for both intercept and slope, while we take the two-dimensional identity matrix for the initial state covariance. The transition matrices are also given by the two-dimensional identity matrix. And the result means are our desired time-varying deltas or hedge ratios.

*3.2 Volatility Proxy*

For the hedge ratio approximation, we also implement volatility proxy to estimate the volatility of the indexes. The Parkinson measure was first proposed by C. Parkinson (1980) as an alternative to the commonly used close-to-close squared return method. Let $(S_1, S_2, \dots, S_{n+j})$ be a set of indexes. Then $V$, the variance of the rate of return on the stock, is traditionally estimated as follows: Letting $r_i = \ln\left(\frac{S_{i+1}}{S_i}\right), i = 1, 2, \dots . n$ = rate of return over the $i^{th}$ time interval, then $V = \frac{1}{n}\sum_i^n r_i^2$. Since we are using the daily based data, our volatility proxy is calculated as $\sigma = \ln(High/Low)$.

In our model, we use the AR (1) model to make prediction of the volatility and then use the predicted volatility to calculate tomorrow's hedge ratio. And since the correlation becomes more and more accurate when the sample size increases, we use $corr_t$ as the estimator of the correlation. Let $\sigma_{t,i}$ be the sqrt of volatility proxy of the index i at time t. We have the hedge ratio between index i and j be $hedge\ ratio_t = \frac{\sigma_{t+1,i}}{\sigma_{t+1,j}} * corr(r_{t,i}, r_{t,j})$. The extreme value method for

estimating the variance of the rate of return assumes that the range of prices reflects the volatility of returns, and that the volatility is proportional to the square root of time. This method is particularly useful for estimating intraday volatility because it is less affected by the bid-ask bounce prices and the overnight price gap, which will affect other volatility estimates.

*4. Copulas*

*4.1 Motivation and Copula*

After picking our pairs to trade, we want to find a way to reveal the relationship in distribution between the two indices. It is difficult to obtain their margin distributions and corresponding parameters. We have to seek an approach to modeling the relationship between two random variables. Copula would be a perfect solution.

Copula is a concept used to describe how several random variables are correlated even though the specific margin distributions of those random variables are unknown. Before defining copula, we introduce the foundation theorem of it.

**Sklar's Theorem (Bivariate):** For two random variables $X_1, X_2$, their univariate margin CDFs $F_1, F_2$ and joint CDF $F$. There exists a copula $C$ such that:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$$

If $X_1, X_2$ are continuous, then such a copula $C$ is unique.

The Sklar's Theorem gives us confidence in its uniqueness when we determine our pair's copula. We do not need to worry about other possible copulas over our sample data after fitting one.

By extracting the empirical function of the daily returns from two indices' historical data, we can map the two unknown return distribution to two U [0,1] distribution $\boldsymbol{U_1}, \boldsymbol{U_2}$. Then we can define the bivariate copula $C$ $[0,1] \times [0,1] \rightarrow [0,1]$ as:

$$C(u_1, u_2) = P(U_1 < u_1, U_2 < u_2) = P\left(X_1 < F_1^{-1}(u_1), X_2 < F_2^{-1}(u_2)\right)$$

By definition, we are able to model the pairs' joint quantile distribution by detecting their own quantile functions. This property rules out the uncertainty of each index's margin distribution.

Considering the mathematical completeness, below we give the formal definition of Copula.

***Definition of Copula:***

If a function $C$ $[0,1] \times [0,1] \rightarrow [0,1]$ such that:

$C(0, u_2) = C(u_1, 0) = 0$ and $C(1, u_2) = u_2$, $C(u_1, 1) = u_1$ for all $u_1, u_2 \in [0,1]$

(2-increasing) For all $u_1^1, u_1^2$ and $u_2^1, u_2^2 \in [0,1]$ with $u_1^1 \leq u_1^2$ and $u_2^1 \leq u_2^2$,

$$C(u_1^2, u_2^2) - C(u_1^2, u_2^1) - C(u_1^1, u_2^2) + C(u_1^1, u_2^1) \geq 0$$

Since Copula is a bivariate joint cumulative distribution function, we can understand its two properties easily. Also, we can derive the conditional distribution of each random variable by taking first derivatives.

$$\frac{\partial C(u_1, u_2)}{\partial u_1} = \frac{\partial P(U_1 < u_1, U_2 < u_2)}{\partial u_1} = P(U_2 < u_2 | U_1 = u_1)$$

$$\frac{\partial C(u_1, u_2)}{\partial u_2} = \frac{\partial P(U_1 < u_1, U_2 < u_2)}{\partial u_2} = P(U_1 < u_1 | U_2 = u_2)$$

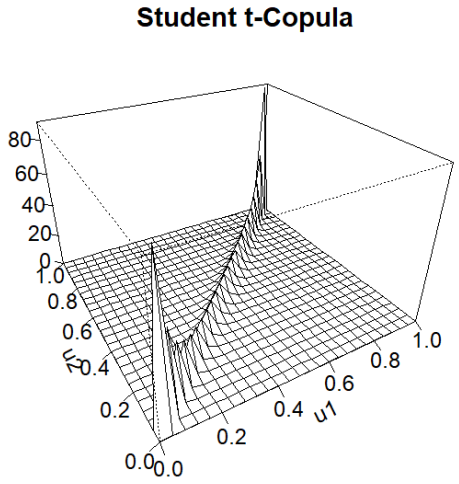
By taking second derivatives with respect to both variables, we have the copula density $c(u_1, u_2)$:

$$c(u_1, u_2) = \frac{\partial C(u_1, u_2)}{\partial u_1 \partial u_2}$$

We can simplify its meaning as the joint quantile density function in pairs trading. Copulas are mainly divided into two categories, Archimedean Copulas and Elliptical Copulas. By looking at their figure of density function, we can see how these different copulas describe the tail dependence of the pair's joint distribution. For example, Frank copula density are relatively high around (0,0) and (1,1), which means there are significant co-movements of the two indices on both upper tail and lower tail, whereas there is only a lower tail peak on the Clayton copula density, which reveals the tendency of going down together. The conditional probability and capacity of measuring tail dependency were the two vital reasons for choosing Copula as our pair trading strategy.

*4.2 Pick the Right Copula*

We use the VineCopula package in R to fit the daily return data train set into an appropriate Copula for our trading pairs. For instance, the chosen pair (IJH - MDY) is fitted in a student-t Copula (Fig xx). On the picture we can see the two indices have significant tail dependence, both upper and lower.



*4.3 Generate Entrance and Exit Signals*

Based on the conditional probability of Copula, we can detect the probability of the return of one leg lower than its current level given the other leg's return. If this probability of one leg is lower than a given threshold, we consider it as underpriced, and it is a signal that implies the stock is going up. Then we can short this leg. Conversely, if the probability is too high beyond an up threshold, we consider its price is overpriced. Then we short this leg. Strategies are same to the other legs. In order to ensure that an entrance opportunity is valid sufficiently, we decide to open a position when double-sided signals emerge.

Here is our entrance signal generating rule:

1) If $\frac{\partial C(u_1,u_2)}{\partial u_1} = P(U_2 < u_2|U_1 = u_1) \leq b_{lo}$ and $\frac{\partial C(u_1,u_2)}{\partial u_2} = P(U_1 < u_1|U_2 = u_2) \geq b_{hi}$,

we long leg 1 and short leg 2 along with the hedge ratio.

2) If $\frac{\partial C(u_1,u_2)}{\partial u_1} = P(U_2 < u_2|U_1 = u_1) \geq b_{hi}$ and $\frac{\partial C(u_1,u_2)}{\partial u_2} = P(U_1 < u_1|U_2 = u_2) \leq b_{lo}$,

we short leg 1 and long leg 2 along with the hedge ratio.

We regard the first time (first day in our test) that the entrance signal does not appear after we already have a position, which means returns of two legs reverts to the normal value (i.e., both two indices are at the corresponding statistical level revealed by their copula) as our exit signal. Then we exit our positions, waiting for next entrance signal.

Below are the rules for continuous signals:

1) If an entrance signal appears after an entrance signal, we update a new position to the old one.

2) If an exit signal appears after an entrance signal, clear all position.

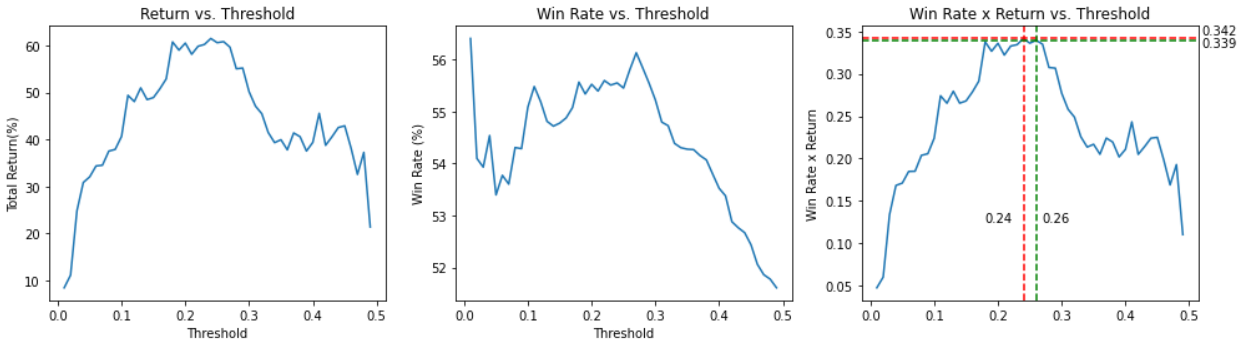
3) If an opposite direction signal (for example, long leg 1then short leg 1) appears, clear all previous position and simply open position following the newest signal (short leg 1).

Then we need to determine the hedge ratio between two legs.

*5. Back-testing*

**Pair 1: IJH - MDY**

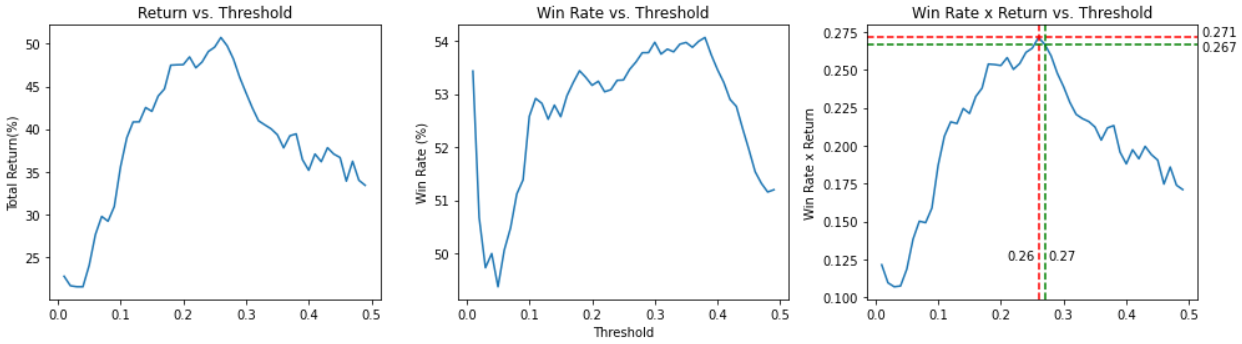
**In-sample performance of Volatility Proxy hedge ratio**




Ideally, as the threshold increases, the win rate decreases. Because the more unlikely a deviation occurs, the more likely the pairs are relatively mispriced. As this trend is visible in the above win rate plot, we think it is fair to say that the copula model gives trading signals better than pure guessing. However, the win rate curve is bumpy in the range of 0.01 to 0.3. When the threshold is around 0.25, the win rate is particularly high.

Nevertheless, we find the sweet spot that gives us the best win rate-return balance is around 0.25.

**In-sample performance of Kalman filter hedge ratio**

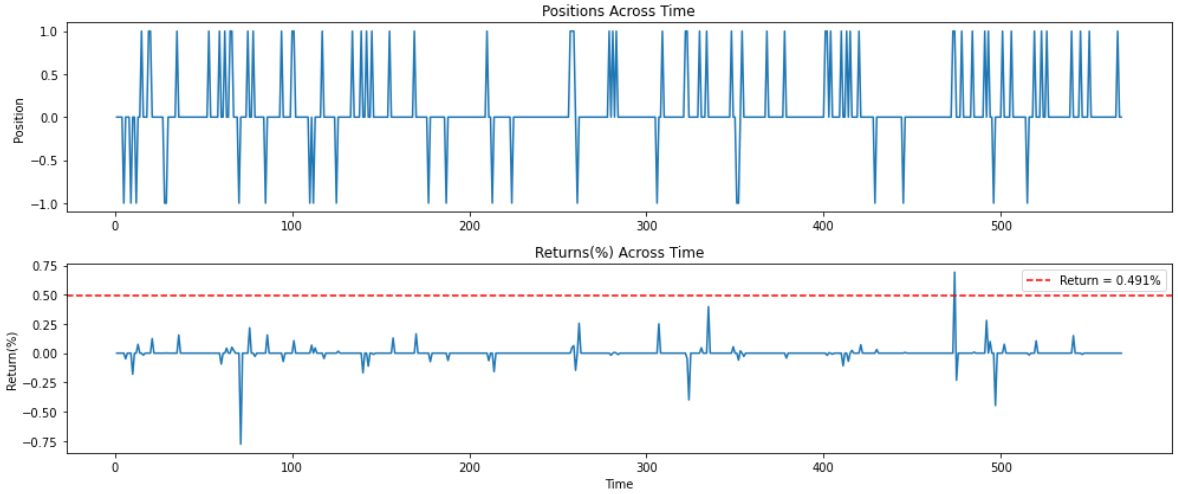



The win rate curve has a similar shape as the one produced by the volatility proxy hedge ratio. Starting from 0.38, the win rate sharply declines. Similar to the volatility proxy, the win rate between 0.2 and 0.3 is higher than what we expected. We offer a few potential explanations: (i) Copula is imperfect. It is a purely statistical model that reflects the long-term distribution of returns. It cannot predict microstructure or react to short-term events. (ii) Pure randomness could also be a reason. It could so happen that over the twenty years, using 0.05 as the threshold gives you a lower probability of profit.

We also offer a few possible explanations for the optimal hedge ratio of 0.25: (i) The composition of the index pairs matters. IJH and MDY are both Russell 1000 mid-size indexes, so they should co-move closely, which enables a relatively loose threshold to provide meaningful signals of mispricing. (ii) If we think of the index prices as a process with a drift (which could be

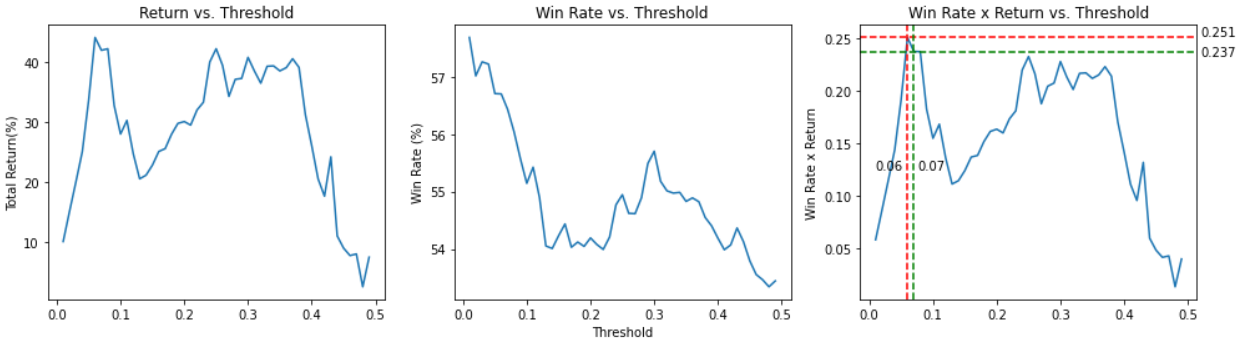
compensation for factors like inflation), then taking a market-neutral position is earning us the drift. Therefore, the more frequently we enter a position, the more drift we earn.

### Out-of-Sample Performance

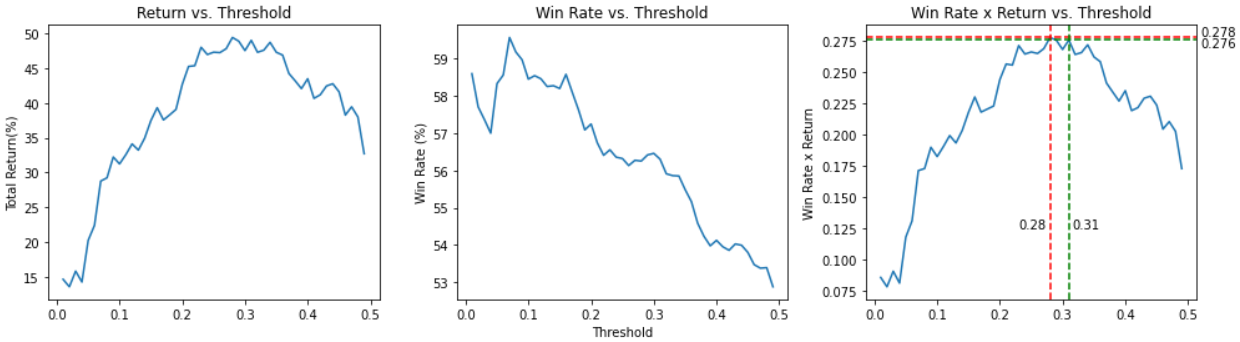



## Pair 2: SPT-VTI

### T-Student Copula + Volatility Proxy Hedge Ratio Performance




### T-Student Copula + Kalman Filter Hedge Ratio Performance



Besides the IJH-MDY pair, we have also found the SPY-VTI pair. Since the copula we fit is always student t copula, here we do not perform the train-test split, instead, show the performance using all the data.

Here we could clearly see the win rate decreases as the threshold increases, and bumps in the win rate when the threshold is around 0.3(in IJH-MDY it was 0.25). This further shows the copulas model is better than guessing. It is also worth noting that the Kalman Filter hedge ratio performs better for this pair and provides a more ideal win rate curve. More assumptions about the index are needed to determine what method is superior, but our opinion is that if one seeks convenience, the volatility proxy method is better, as the computation is very simple, and its hedge ratio could be easily explained. If one seeks more non-linearity, the Kalman filter is a better way as it uses conditional probability instead of OLS to estimate the ratio.

*6. Summary*

Based on the above research, we draw the following conclusions:

1. Choosing a good pair is important to start with. A pair with an unstable relationship in returns makes estimating the hedge ratio difficult and inaccurate, resulting in more risk losing.

2. Copula provides valuable insights into abnormal in relative returns between a good pair. When using copula to trade a pair with a stable relationship, the hedge ratio becomes less important as we can easily find a fairly good estimate of the hedge ratio.

3. If one assumes that the market is efficient (like us), i.e., the market will correct itself quickly, one should set the threshold to keep a position strictly. The benefit of doing so is one could avoid unexpected permanent changes in the relationship between two indexes. Copula cannot quickly detect a structural change or a shock as it fits in large data sets, any new data affects its structure with little to no effect. With such a robust model, it is important to design protect mechanism in the system or keep a close eye on market movements by humans.

# References

Copula for pairs trading: A detailed, but practical introduction. Hudson & Thames. (2023, February 5). Retrieved February 28, 2023, from https://hudsonthames.org/copula-for-pairs-trading-introduction/

Dynamic hedge ratio between ETF pairs using the Kalman filter. QuantStart. (n.d.). Retrieved February 27, 2023, from https://www.quantstart.com/articles/Dynamic-Hedge-Ratio-Between-ETF-Pairs-Using-the-Kalman-Filter/

Liew, R.Q. and Wu, Y., 2013. Pairs trading: A copula approach. Journal of Derivatives & Hedge Funds, 19(1), pp.12-30.

Nelsen, R.B., 2007. An introduction to copulas. Springer Science & Business Media.

Parkinson, Michael. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. The Journal of Business. 53. 61-65. 10.1086/296071.

PARTIAL KENDALLS TAU CORRELATION：
https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/partktau.htm

Performance of a Relative-Value Arbitrage Rule, The Review of Financial Studies, Volume 19, Issue 3, Fall 2006, Pages 797–827, https://doi.org/10.1093/rfs/hhj020

State space models and the Kalman filter. QuantStart. (n.d.). Retrieved February 27, 2023, from https://www.quantstart.com/articles/State-Space-Models-and-the-Kalman-Filter/

Xie, W., Liew, R.Q., Wu, Y. and Zou, X., 2016. Pairs trading with copulas. The Journal of Trading, 11(3), pp.41-52.