NUS

National University
of Singapore

Risk Management Institute

MSc in Financial Engineering

Statistical Arbitrage: Integrating Machine Learning Clustering with Stochastic Modelling

FE5221 - Trading Principles and Fundamentals

Hui Jia Shun

A0250406R

November 2023

View formation

Pairs trading, a subset of statistical arbitrage strategies, capitalizes on the quantifiable relationships between assets. Predominantly employed by hedge funds, this approach hinges on the principle that cointegrated pairs of assets, often equities, tend to revert to their historical spread average. This mean-reversion characteristic forms the cornerstone of pairs trading. When two cointegrated stocks diverge, a trade is initiated by shorting the overperforming stock and going long the underperforming one. The trade is then closed when the spread reverts to its mean, capturing the profit of convergence. This strategy is rooted in relative value as it assumes that the prices of two stocks will eventually move back to historical equilibrium, profiting from temporary inefficiencies.

However, this is a common strategy in hedge funds and an overcrowding of trades relying on this strategy will eventually cause the alpha to decay, reducing the potential for outsized returns. Recognizing this challenge, we explore innovative ways to generate additional value and maintain the viability of this investment strategy in a competitive landscape. We approach this problem by clustering companies by their fundamental similarities instead of the traditional GICs framework.

Trade Selection

A common approach in statistical arbitrage involves grouping securities based on shared attributes, making it more logical and defensible to assume a stable relationship between them. Most players typically use industry classifications to narrow down their choices for trading. Our approach focuses on the S&P 500, examining three different clustering techniques as potential methods. These include agglomerative hierarchical clustering, K-Means clustering, and the conventional GICs grouping. The unsupervised clustering algorithms (agglomerative hierarchical clustering and K-Means) are trained on fundamental factor exposures over a rolling 60-day timeframe. This is done to detect short-term resemblances among the stocks as we are assuming similarity between stocks evolves with time. Having a 60-day window in running the clustering tests seeks to capture that specific dynamic. The parameterization and selection of the number of clusters will follow the GICs framework and set at 70.

Asset Name	GICS Industry	K-Means Cluster
EBAY INC	Internet & Direct Marketing Retail	5
AMAZON.COM INC	Internet & Direct Marketing Retail	54
ETSY INC	Internet & Direct Marketing Retail	0
ORACLE CORP	Software & Services	1
MICROSOFT CORP	Software & Services	1
CADENCE DESIGN SYSTEMS INC	Software & Services	1
SYNOPSYS INC	Software & Services	1

To confirm the statistical connections between securities, a variety of tests are applied. In this approach, we utilize the Engle-Granger cointegration test to evaluate the cointegration of pairs within each cluster over the rolling previous 60 days, like the clustering steps. Next, we use the Ornstein-Uhlenbeck process to model the spreads, calibrating the model based on the same 60-day historical data. This calibration helps us determine the mean and standard deviation of the spreads. On the day of trading, we use the z-score calculated from the mean and standard deviation obtained through the OU process to guide our trading decisions. Our strategy involves entering a trade when the z-score exceeds an absolute value of 1 (representing one standard deviation), where we short the outperforming stock and go long on the underperforming one, and exiting the trade when the z-score falls to 0.5. Additionally, a stop loss is set at a z-score of 2 to capture potential risk where the statistical relationship breaks down.

Capital Allocation

In the realm of capital allocation, we adopted a risk-adjusted approach. With an AUM of one million dollars, each trade was sized at USD $ 10,000, allowing for a diversified portfolio of pair trades while maintaining a manageable exposure level to individual positions. In terms of sizing the position of the pairs, we determine the hedge ratio using the security's CAPM beta. Our assumption is that the strategy has idiosyncratic alpha agnostic to the market, and we would like to reduce market risk by isolating the specific, desired source of return. Beta hedging can also bring down overall portfolio volatility, improving the Sharpe ratio. Also, the strategy seeks to exploit short-term price inefficiencies and beta hedging helps to ensure that the returns are driven by inefficiencies instead of broad market movements.

Buy	Sell	Buy Size	Sell Size
YUM BRANDS INC	CINTAS CORP	5375.96	4624.03
CORTEVA INC	ELECTRONIC ARTS INC	5299.40	4700.59

Execution Strategy

When initiating a trade, it's crucial to consider both explicit and implicit transaction costs. This requires a high degree of confidence in the signal before executing a trade, as those with low conviction are not pursued. Due to the strategy's reliance on short-term signals and the rapid decay of alpha over time, prompt execution of trades is essential. Thus, using a market order is recommended for trade execution. With the fund's size at one million, the market impact is likely minimal, especially considering the S&P 500's vast market size of approximately 30 trillion, making it a highly liquid market.

Monitor and Review

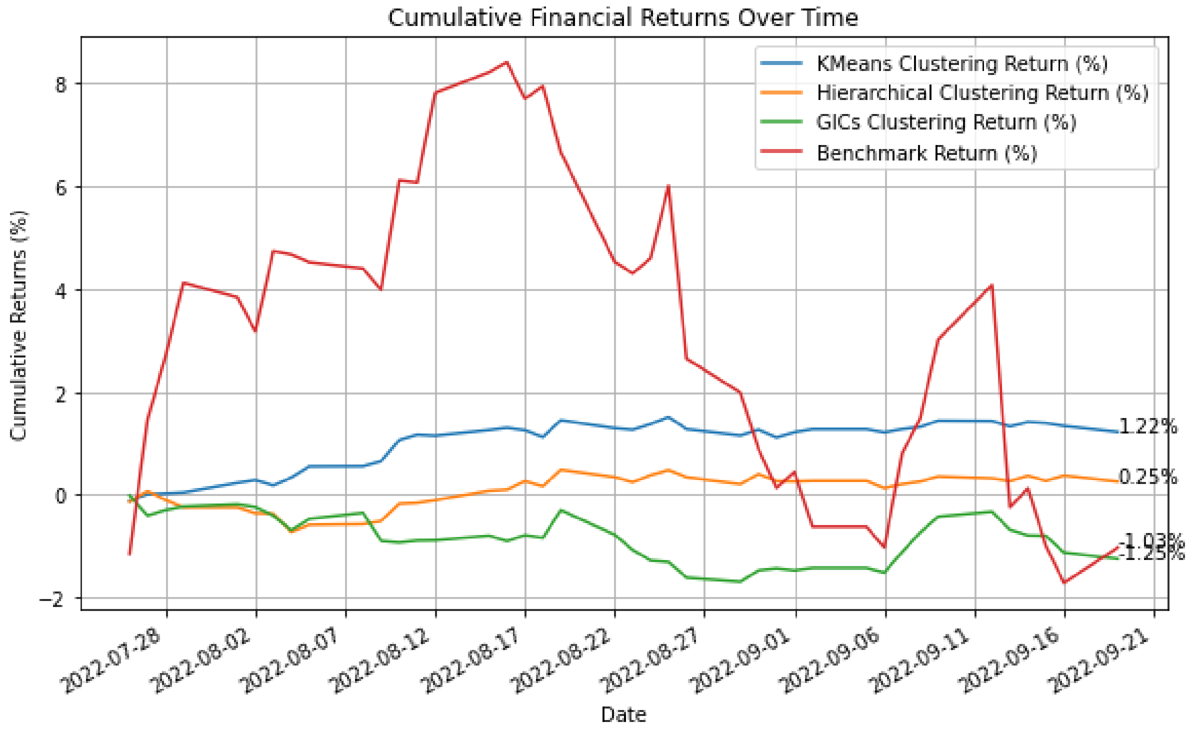
In evaluating the performance of the strategy, we leverage on S&P 500 as the benchmark. The back test was conducted over the interval from July 26, 2022, to September 19, 2022. It revealed that the approach employing K-Means for clustering securities surpassed the alternative two clustering methods. It achieved a 1.22% performance increase, compared to 0.25% for hierarchical clustering and a -1.25% result for GICs. Additionally, implementing a beta hedge reduced volatility, thereby enhancing the Sharpe ratios. This led to better risk-adjusted returns for all three strategies. The realized beta is low (0.028 – 0.044), demonstrating the success of our hedging strategy, which is based on a predictive ex-ante model forecast. The K-Means strategy also exhibited the smallest maximum drawdown. This suggests that the strategy not only mitigates risk but also yields appealing returns.

Portfolio Name	**Portfolio Return (%)**	**Portfolio Volatility (%)**	**Sharpe Ratio**	**Maximum Drawdown (%)**
StatArb Hierarchical Clustering	0.2502	0.9057	-0.110884	-0.7917
StatArb K-Means	1.2231	0.8323	1.048212	-0.3958
StatArb GICs	-1.2522	1.4723	-1.088660	-1.6877

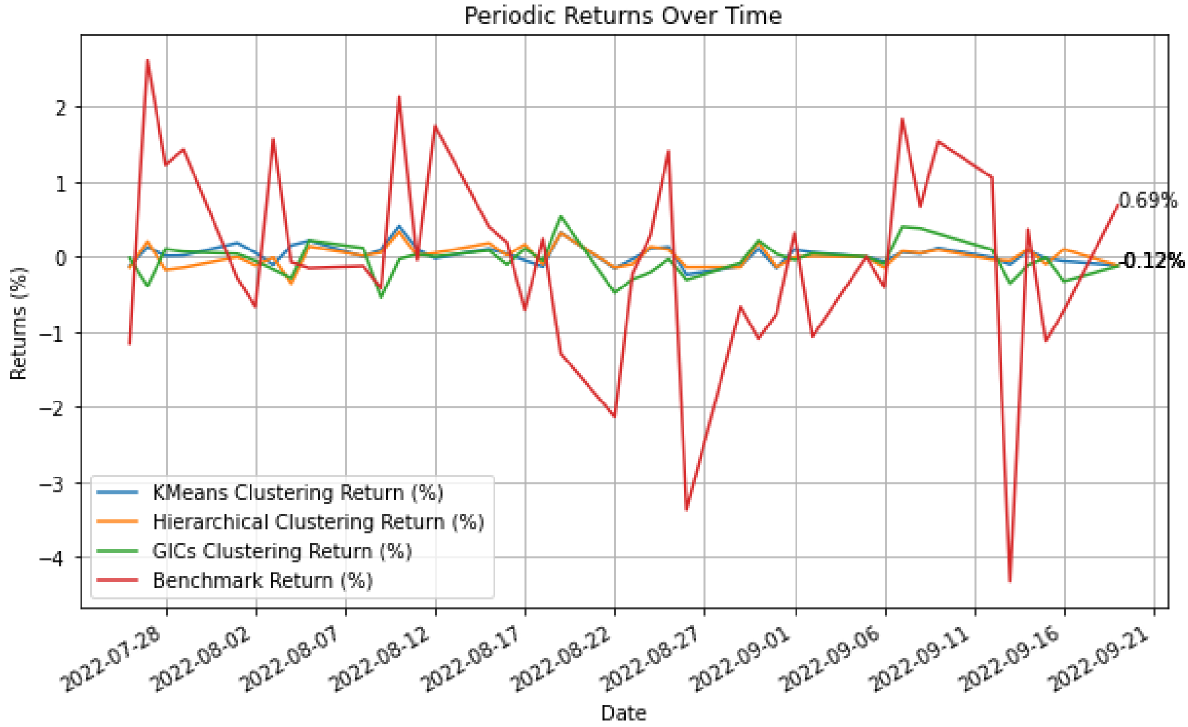
Regarding enhancements to the strategy, several possibilities exist. Firstly, we can utilize parameters derived from the OU process to simulate the first passage time, predicting the expected time required to reach profit targets. Subsequently, we will terminate the trade if the spreads converge slower than the expected threshold.

Another improvement involves adapting to the current high-speed trading environment, which is crucial for leveraging short-term price anomalies. This necessitates integrating high-frequency data into our strategy implementation. Nonetheless, adopting such an approach demands a considerable investment in developing the necessary infrastructure.

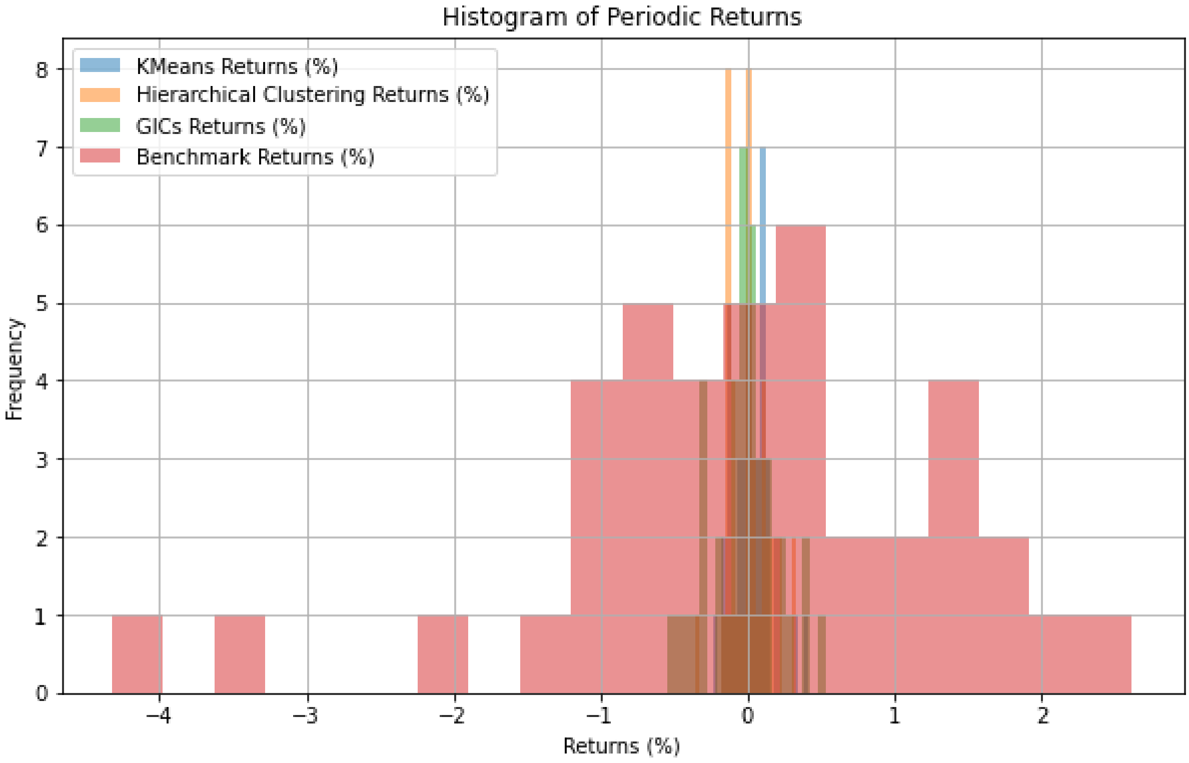
Appendix



Cumulative Financial Returns Over Time
KMeans Clustering Return (%)
Hierarchical Clustering Return (%)
GICs Clustering Return (%)
Benchmark Return (%)
Cumulative Returns (%)
8
6
4
2
0
−2
1.22%
0.25%
-1.03%
-1.25%
2022-07-28
2022-08-02
2022-08-07
2022-08-12
2022-08-17
2022-08-22
2022-08-27
2022-09-01
2022-09-06
2022-09-11
2022-09-16
2022-09-21
Date



Periodic Returns Over Time
Returns (%)
2
1
0
−1
−2
−3
−4
0.69%
-0.12%
KMeans Clustering Return (%)
Hierarchical Clustering Return (%)
GICs Clustering Return (%)
Benchmark Return (%)
2022-07-28
2022-08-02
2022-08-07
2022-08-12
2022-08-17
2022-08-22
2022-08-27
2022-09-01
2022-09-06
2022-09-11
2022-09-16
2022-09-21
Date


Histogram of Periodic Returns
KMeans Returns (%)
Hierarchical Clustering Returns (%)
GICs Returns (%)
Benchmark Returns (%)
Frequency
0
1
2
3
4
5
6
7
8
−4
−3
−2
−1
0
1
2
Returns (%)

Portfolio Name	Benchmark Name	Start Date	End Date	Portfolio Return (%)	Benchmark Return (%)	Active Return (%)	Portfolio Volatility (%)	Benchmark Volatility (%)	Beta
StatArb Hierarchical Clustering	S&P 500 Index	2022-07-25	2022-09-19	0.2502	-1.3976	1.6478	0.9057	8.7628	0.028680
StatArb K-Means	S&P 500 Index	2022-07-25	2022-09-19	1.2231	-1.3976	2.6207	0.8323	8.7628	0.038497
StatArb GICs	S&P 500 Index	2022-07-25	2022-09-19	-1.2522	-1.3976	0.1454	1.4723	8.7628	0.044025

Portfolio Name	Sharpe Ratio	Sortino Ratio	Treynor Ratio	Alpha (%)	Jensen's Alpha (%)	Tracking Error (%)	Information Ratio	Upside Capture Ratio (%)	Downside Capture Ratio (%)	Maximum Drawdown (%)
StatArb Hierarchical Clustering	-0.110884	-0.158765	-0.035019	0.2903	-0.0503	8.5559	0.192593	3.5980	2.7674	-0.7917
StatArb K-Means	1.048212	1.827287	0.226621	1.2769	0.9397	8.4598	0.309780	4.6741	-1.1093	-0.3958
StatArb GICs	-1.088660	-1.350436	-0.364082	-1.1907	-1.5259	8.4967	0.017107	0.9420	7.6751	-1.6877