Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2022.45.3.123

Vol. 45, No. 3 : 123-130, September 2022
ISSN : 2005-0461 / EISSN : 2287-7975

Clustering-driven Pair Trading Portfolio Investment in Korean Stock Market

Poongjin Cho* · Minhyuk Lee** · Jae Wook Song*†

*Department of Industrial Engineering, Hanyang University
**Department of Business Administration, Pusan National University

한국 주식시장에서의 군집화 기반 페어트레이딩 포트폴리오 투자 연구

조풍진* · 이민혁** · 송재욱*†

*한양대학교 산업공학과
**부산대학교 경영학부

Pair trading is a statistical arbitrage investment strategy. Traditionally, cointegration has been utilized in the pair exploring step to discover a pair with a similar price movement. Recently, the clustering analysis has attracted many researchers' attention, replacing the cointegration method. This study tests a clustering-driven pair trading investment strategy in the Korean stock market. If a pair detected through clustering has a large spread during the spread exploring period, the pair is included in the portfolio for backtesting. The profitability of the clustering-driven pair trading strategies is investigated based on various profitability measures such as the distribution of returns, cumulative returns, profitability by period, and sensitivity analysis on different parameters. The backtesting results show that the pair trading investment strategy is valid in the Korean stock market. More interestingly, the clustering-driven portfolio investments show higher performance compared to benchmarks. Note that the hierarchical clustering shows the best portfolio performance.

Keywords : Pair trading, Clustering, Statistical Arbitrage, Investment Strategy

1. 서 론

페어 트레이딩은 유사한 가격흐름을 지닌 두 종목 간의 스프레드를 활용하는 투자전략으로 주로 기업의 펀더멘털보다는 주식 시계열 특성에 기반 한 통계적 차익거래(Statistical Arbitrage)에 집중한다. 비슷한 가격 움직임을 가졌던 종목 간의 가격 차가 통계적 범위를 벗어났을 때, 다시 평균으로 회귀한다는 가정 하에 스프레드 축소에 베팅한다. 물론, 장기적 균형관계를 가지던 두 종목 간 스프레드가 미래에 발산할 수도 있으며 이는 손실을 야기한다.

페어 트레이딩의 가장 중요한 절차는 비슷한 가격 움직임을 가지는 페어를 탐색하는 단계이다. 이에 따라 여러 연구들이 오래전부터 진행되어왔다. 전통적인 방법은 공적분(Co-integration)의 활용이다. 두 시계열 간 공적분 관계의 존재는 장기적 균형관계를 의미하며, 두 시계열 간 안정관계가 깨지더라도 이 상태가 장기간 지속되지 않고 이전의 안정적인 관계로 회귀하는 성질을 가진다. 따라서 평균 회귀 형태의 Dickey-Fuller 검정을 통해 유사한 가격흐름을 지닌 페어를 탐색할 수 있다[7]. 그러나, 헤지펀드 간 경쟁이

Received 22 August 2022; Finally Revised 15 September 2022; Accepted 16 September 2022
† Corresponding Author : jwsong@hanyang.ac.kr

심화됨에 따라 단순한 페어 탐색으로 창출 가능 한 수익은 꾸준히 기회가 줄어들어왔다. 최근에는 이를 해결하고자 기계학습을 통한 페어 탐색 연구들이 많이 진행되었다. Han et al.[11]는 페어를 탐색하는 과정에서 비지도 학습 중 하나인 군집화 방법 적용을 제안하였으며, 미국 주식 시장인 NYSE, AMEX, Nasdaq에 속한 종목에 대해 군집화 기반 페어트레이딩 전략의 수익성을 입증하였다. 군집화 방법은 페어 트레이딩 뿐만 아니라 포트폴리오 최적화, 리스크 관리 등 금융 분야에서 많이 사용되어 왔다. 본 연구에서는 Han et al.[11]에서 제시 된 방법론에 기초하여 한국 주식 시장에서도 군집화를 통한 페어 트레이딩 전략의 수익성이 존재하는지 확인하였다. 코스피 지수에 속한 종목 중 비슷한 주가 흐름을 가지는 주식을 포착하고, 군집화가 잘 되었는지 확인하고자 페어 트레이딩 전략을 통해 수익성을 평가하였다.

페어 트레이딩의 자세한 과정은 다음과 같다. 먼저, 순차적 단계를 1) 유사도 포착, 2) 스프레드 포착, 3) 수익률 테스트 3개로 나눈다. 유사도 포착 단계에서는 비슷한 가격 움직임을 갖는 주식끼리 군집화를 진행하고, 스프레드 포착 단계에서는 각 군집에서 가장 큰 스프레드를 가지는 페어를 찾아서 포트폴리오에 해당 페어를 포함시킨다. 수익률 테스트 단계에서는 포트폴리오에 포함시킨 페어들에 대해서 저평가 종목은 매수를, 고평가 종목은 공매도를 하는 롱숏전략을 수행한다. 포트폴리오에 여러 페어들이 포함되었다면 동일비중으로 롱숏전략을 진행하고, 해당 과정은 주초의 시가에 시작하여 주말의 종가에 마무리하는 것을 매주 반복한다. 코스피 지수에 속한 종목의 주식 가격 데이터만을 사용하였고, 2000년 1월부터 2020년 12월까지의 데이터를 한국거래소를 통해 얻었다. 가격흐름을 측정하는 방법으로는 주가의 로그 수익률을 이용한 주가 모멘텀을 사용하였고, 최소-최대 정규화 과정을 통해 0과 1 사이의 수로 변환하였다. 군집화 방법으로는 K-평균(K-means) 알고리즘과 병합계층(Agglomerative Hierarchical) 군집 알고리즘을 사용하였다. 이 때, 군집화 방법 내 거리 함수로는 L1인 맨하탄(Manhattan)과 L2인 유클리드(Euclide)를 사용하였다. 페어 트레이딩 전략의 수익성은 다양한 포트폴리오 지표를 통해 검증하였다. 벤치마크로는 코스피 지수의 매수 후 보유 전략과 공적분을 통한 페어 트레이딩 전략을 인용하였고, 추가적으로 기간별 분석과 민감도 분석을 통해 백테스팅 결과의 강건성을 뒷받침하였다. 이어지는 제2장에서는 선행연구와 이론적 배경에 대해 설명하고, 제3장에서는 군집화 기반 페어 트레이딩 투자전략 방법론을 소개한다. 제4장에서는 실험을 통해 백테스팅의 수익성을 살펴본다. 마지막으로 제5장에서는 결론을 도출하고 추후 연구과제들을 제시한다.

2. 이론적 배경

2.1 선행연구

비슷한 주가 움직임을 지닌 페어를 탐색하는데 있어 초기에는 공적분이 많이 사용되었고[5], Wen et al.[24]는 공적분 행렬의 최소 신장 트리를 구축하여 네트워크 관점으로 페어 트레이딩의 특징을 분석하였다. 또한, Ramos-Requena et al.[19]는 탐색된 페어에 대해서 Hurst 지수와 공적분을 결합하여 최적의 포트폴리오 비중을 결정하는 모델을 개발하였고, Flori and Regoli[8]는 공적분을 통해 탐색된 페어에 대해 LSTM을 통해 주가가 상승할 확률을 계산하였다. 두 시계열 데이터 간의 연관도를 측정하는데 있어서 공적분을 사용하지 않는 다른 시도들도 수행되어 왔다. Gupta and Chatterjee[10]는 동적 교차상관을 이용한 거리 함수를 제안하였다. 또한, 혼합정수계획법을 통해 다중 목적함수 유전자 알고리즘을 풀어서 페어를 탐색하려는 시도도 존재하였으며[9], Sarmento and Horta[22]는 주성분분석과 OPTICS(Ordering Points to Identify the Clustering Structure)를 결합한 페어 탐색 기법을 개발하였다.

2.2 K-평균 군집

K-평균 군집은 비지도학습 중 가장 널리 알려진 방법이다[12]. $X=\{x_1, x_2, ..., x_m\}$를 데이터 포인트의 집합이라 하고, $C=\{c_1, c_2, ..., c_n\}$를 각 군집의 센트로이드(Centroid)로 이루어진 집합이라고 할 때, 벡터 $W=[w_{ij}]_{m\times n}$에 대한 원소 w_{ij}를 다음과 같이 정의할 수 있다.

$$w_{ij}=\begin{cases}1\,(x_i\in clusterk)\\0\,(x_i\notin clusterk)\end{cases} \tag{1}$$

이 때, K−평균 군집의 목적함수는 다음과 같다.

$$O(W,C)=\sum_{i=1}^{m}\sum_{j=1}^{n}w_{ij}\|x_i-c_j\|^2 \tag{2}$$

초기값 c_j를 설정한 뒤, 다음의 2가지 과정이 반복된다. 첫째로, 고정된 c_j에 대해서 목적함수 $O(W, C)$를 w_{ij}에 대해 최소화하고, 이는 각 데이터 포인트를 근접한 군집에 할당하는 과정이다. 둘째로, 고정된 w_{ij}에 대해서 목적함수 $O(W, C)$를 c_j에 대해 최소화하고, 이는 각 군집의 센트로이드를 재설정하는 과정이다. 이 2가지 과정을 계속해서 반복한 뒤, 데이터 포인트의 할당된 군집이 변하지 않거나 초기 설정한 최대 반복횟수에 도달하면

반복을 중지한다. 초기값을 설정하는 데에는 Arthur and Vassilvitskii[4]가 제시한 K-평균++ 알고리즘을 사용하였다.

K-평균 군집은 모든 데이터 포인트를 각 군집에 할당하기 때문에 이상치를 포착해내는 능력이 없다. 이러한 이상치는 투자전략의 수익성을 방해할 수 있다. 따라서 K-평균 군집 알고리즘을 수행한 뒤 이상치를 제거하는 Hautamäki et al.[13]의 방법을 해당 연구에서는 채택하였고, 자세한 과정은 다음과 같다. 각 데이터 포인트에 대해, 최근접 이웃까지의 거리를 구하고 오름차순으로 정렬하여 α 백분위수가 되는 거리를 이상치 임계값 ϵ라고 하자. 센트로이드에서 각 데이터 포인트에 대한 거리를 측정하여 그 거리가 이상치 임계값 ϵ보다 크다면, 그 데이터 포인트는 이상치로 간주한다. 이 때, 거리를 측정할 때 사용하는 거리 함수는 L2를 사용하였고, 하이퍼파라미터는 군집의 개수와 이상치 임계값이다.

2.3 계층적 군집

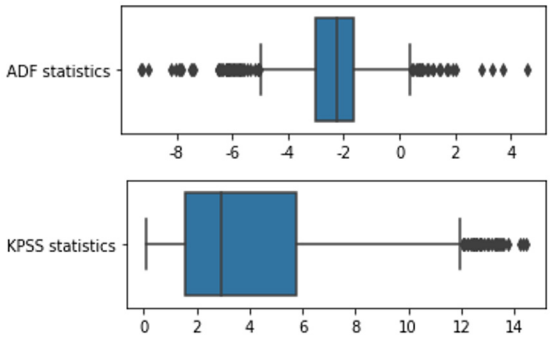
계층적 군집은 병합 계층 군집과 분할 계층 군집으로 나뉜다[14]. 해당 연구에서는 하나의 군집이 남을 때까지 계속해서 가장 가까운 군집을 합치는 병합 계층 군집[6]을 사용하였으며 자세한 과정은 다음과 같다. 먼저, 모든 데이터 포인트끼리 거리 행렬을 계산하고, 각 데이터 포인트를 단일 군집으로 간주한다. 다음으로 가장 가까운 두 군집을 합치는데, 이 때 거리는 군집에 속한 모든 구성 요소들끼리 평균 거리를 채택하였다[17]. 두 군집이 합쳐짐에 따라 모든 군집끼리 거리 행렬을 계산할 수 있고, 하나의 군집이 남을 때까지 위의 과정을 반복한다.

K-평균 군집 알고리즘과 마찬가지로 계층적 군집 알고리즘에서도 이상치를 다음과 같이 포착할 수 있다. 각 군집끼리 거리 행렬을 계산하고 거리에 대해 오름차순으로 정렬하여 α 백분위수가 되는 거리를 이상치 임계값 ϵ라고 하자. 두 군집 간의 거리가 이상치 임계값 ϵ보다 크면 두 군집은 병합하지 않는다. 병합 과정을 반복함에도 병합되지 않는 군집은 이상치 군집으로 간주한다. 이 때, 거리를 측정하는데 있어서 L1 거리 함수를 사용하였고[1] 하이퍼파라미터는 군집의 개수 혹은 이상치 임계값이다. 해당 연구에서는 이상치 임계값의 백분위수인 α를 기준으로 하이퍼파라미터 최적화를 진행하였다.

3. 군집화 기반 페어트레이딩 전략

3.1 데이터 전처리

코스피 지수에 속한 종목들에 대해 2000년 1월부터 2020년 12월까지의 가격 데이터를 한국거래소로부터 얻었다. 주가 데이터는 비정상성을 가지므로 직접 사용되기 보다는 정상성을 띄는 수익률이나 가격 모멘텀으로 변환하여 사용한다. 주가 데이터가 비정상성을 띄는지 확인하기 위하여 각 종목에 대해 ADF 통계량과 KPSS 통계량을 구하였고, 이를 통한 박스플롯 결과는 <Figure 1>와 같다.



<Figure 1> Test Statistics of ADF Test and KPSS Test

5% 유의수준에서 ADF 통계량의 임계치는 -2.871이고, KPSS 통계량의 임계치는 0.463이다. 대부분의 주가 데이터가 비정상성을 띄는 것을 확인할 수 있고, 로그 수익률 변환을 통해 주가 데이터의 정상성을 확보할 수 있다. 해당 연구에서는 가격흐름을 측정하기 위하여 가격 모멘텀을 사용하였다. 가격 모멘텀이 비슷한 주식들은 가까운 미래에 비슷한 주가 움직임을 가진다고 가정할 수 있으므로 페어 트레이딩에 사용할 수 있다. 가격 모멘텀을 정의하기에 앞서, t 시점에서 주식 i에 대한 수익률이 $R_i(t)=\ln(P_i(t))-\ln(P_i(t-1))$와 같다면, t 시점에서 n일 동안의 주식 i에 대한 가격 모멘텀 $m_{i,n}(t)$는 다음과 같다.

$$m_{i,n}(t)=\begin{cases} R_i(t) & (n=1) \\ \sum_{s=t-n+1}^{t-1} R_i(s) & (n=2,3,...,60) \end{cases} \quad (3)$$

유사한 주가를 포착하기 위해 과거 60영업일 간의 주가 모멘텀을 사용한다. 군집화에 앞서 가격 모멘텀은 0부터 1까지의 값을 가지도록 최소-최대 정규화 과정을 통해 변환한다.

3.2 군집화 기반 페어 포착 및 투자전략

투자전략은 3가지 단계로 나뉜다. 첫 번째는 유사도 포착 단계로써, 현재를 t로 기준 삼았을 때 $t-65$에서 $t-6$까지의 시계열이 해당된다. 각 종목에 대해 유사도

포착 단계 동안의 60개 주가 모멘텀을 준비하고 군집화를 수행한다. 각 군집은 비슷한 주가 움직임을 가지는 종목끼리 묶이게 된다. 두 번째는 스프레드 포착 단계로써, $t-5$에서 $t-1$까지의 시계열이 해당된다. 각 군집에서 가장 높은 수익률을 가진 종목과 가장 낮은 수익률을 가진 종목을 짝짓는다. 또한, 두 번째로 높은 수익률을 가진 종목과 두 번째로 낮은 수익률을 가진 종목끼리 짝을 짓는다. 마찬가지 방법을 통해 스프레드 포착 구간 동안의 수익률을 정렬하여 각 군집에서 페어를 구성한다. 그 다음, 모든 페어들의 수익률 차이를 구하고, 이 차이들의 표준편차를 계산한다. 만약, 어떤 페어의 스프레드가 이 표준편차를 능가한다면 포트폴리오에 해당 페어를 편입시키고, 반대로 어떤 페어의 스프레드가 이 표준편차를 넘지 못한다면 해당 페어는 투자하지 않는다. 해당 과정을 거친 뒤 포트폴리오는 여러 페어들을 포함하고 있고, 테스트 구간 동안 롱숏전략을 수행하게 된다. 테스트 구간은 t에서 $t+5$까지이고, 테스트 구간의 시작은 1주일 시작일의 시작가에 진행되고, 끝은 1주일 마지막일의 종가에 마무리된다. 두 시계열 간에 유사도가 포착되었다는 것은 장기적 균형 관계가 이루어졌다는 뜻이고, 잠시 안정관계가 깨지더라도 그 상태가 장기간 지속되지 않을 것이다. 따라서, 스프레드 포착 구간과 테스트 구간의 길이는 유사도 포착보다 상대적으로 짧게 설정해야 했고, 1주일 이상 스프레드가 유지된다면 발산할 가능성이 있기 때문에 스프레드 포착 구간과 테스트 구간의 길이는 5영업일로 설정하였다. 유사도 포착 구간의 경우, 스프레드 포착 구간보다는 긴 구간으로 설정하여야 하지만, 너무 긴 구간을 설정할 시 최근 데이터의 중요도가 줄어들기 때문에 1분기로 판단하였고, 60영업일로 설정하였다.

포트폴리오의 수익률을 계산하는 데에는 3가지 과정이 진행된다. 먼저, 탐색된 페어들의 각 수익률을 계산한다. 저평가되었다고 판단한 주식은 매수 및 매도하여 매수 포지션의 수익률을 구하고, 고평가 되었다고 판단한 주식은 공매도 및 청산하여 매도 포지션의 수익률을 구한다. 두 수익률을 빼면 롱숏전략의 수익률이 계산된다. 두 번째로, 각 페어들을 동일비중으로 거래한다고 가정하였으므로 각 페어의 롱숏전략 수익률에 대한 평균을 계산한다. 마지막으로, 해당 수익률은 1주일간의 수익률이므로 수익성 지표를 계산하기 위해서는 연간 수익률로 변환한다.

3.3 포트폴리오 수익성 지표

해당 투자전략의 수익성을 측정하기 위해서 다음의 5가지 대표적인 지표를 사용한다. 샤프 비율(Sharpe ratio, ShR)은 초과수익률을 수익률의 표준편차로 나눈 값으로써 위험 대비 수익률을 의미하고[23], 이 비율이 높다면 수익성이 높다고 평가할 수 있다. 포트폴리오의 하방 위험만으로 표준편차를 계산하여 산출되는 소르티노 비율(Sortino ratio, SoR)은 손실 위험 대비 수익률을 계산한다. 이익 승수(Profit factor, PF)는 총 수익 대비 총 손실을 의미하며 해당 포트폴리오를 통해 얼마나 안전하게 이익을 얻을 수 있는지 확인할 수 있는 지표이다. 최대 손실 낙폭(Maximum Drawdown, MDD)은 총 포트폴리오 기간 동안 최고점으로부터의 최대 손실을 의미하고[15], 칼마 비율(Calmar ratio, CR)은 수익률 평균 대비 최대 손실 낙폭으로서 칼마 비율이 높을수록 포트폴리오는 더 안전하다[15]. $R_i, \sigma_i, \sigma_i^-, P_i, L_i, PV_i, LP_i$가 각각 포트폴리오 i의 수익률, 표준편차, 하방 표준편차, 총 수익, 총 손실, 피크 값, 피크 이후 가장 낮은 값이고, R_{rf}를 무위험수익률라고 할 때, 위의 5가지 지표에 대한 수식은 다음과 같다.

$$ShR_i = \frac{R_i - R_{rf}}{\sigma_i} \tag{4}$$

$$SoR_i = \frac{R_i - R_{rf}}{\sigma_i^-} \tag{5}$$

$$PF_i = \frac{P_i}{L_i} \tag{6}$$

$$MDD_i = \frac{LP_i - PV_i}{PV_i} \times 100 \tag{7}$$

$$CR_i = \frac{R_i}{MDD_i} \tag{8}$$

5가지 대표적인 지표 외에도 추가적으로 수익률의 평균, 수익률의 표준편차, 하방 표준편차, 총 수익, 총 손실, 수익 기간, 손실 기간 등을 측정하여 해당 투자전략의 수익성을 평가하였다.

군집화 기반 페어트레이딩의 수익성에 대한 상대적 비교를 위해 2가지 벤치마크를 다음과 같이 구성한다. 첫 번째 벤치마크는 코스피 지수를 매수 후 보유 방식으로 투자한다. 두 번째 벤치마크는 전통적 페이트레이딩 방법론인 공적분 기반 포트폴리오와의 비교를 수행한다.

마지막으로 기간별로 수익성이 유지되는지, 머신러닝 모델의 하이퍼파라미터에 강건한지 등을 추가로 분석한다. 군집화를 통해 페어를 포착하는데 있어서 모델의 하이퍼파라미터에 따라 결과가 달라질 수 있다. K-평균 군집의 경우, 군집의 수와 이상치 임계값의 하이퍼파라미터를 가진다. 본 연구에서 K-평균 군집에 대한 이상치 임계값은 0.5를 사용하였고, 군집의 수는 5, 10, 25, 50, 100, 200, 400에 대하여 각각 결과를 산출하고, 가장 큰 샤프 비율을 가지는 경우를 최종 포트폴리오로 확정하였

다. 계층적 군집의 하이퍼파라미터는 이상치 임계값이 있고, 0.1, 0.2,…, 0.9에 대하여 각각 결과를 산출하고, 가장 큰 샤프 비율을 가지는 경우를 최종 포트폴리오로 확정하였다.

4. 백테스팅 결과 및 분석

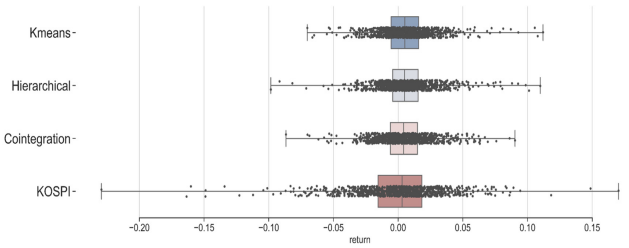
4.1 수익성 평가

해당 투자전략의 수익성이 있는지 평가하고자 2000년 1월부터 2020년 12월까지 백테스팅을 진행하였다. 매주 포착된 페어에 대해 주초 시가에 매수 및 공매도, 주말 종가에 매도 및 상환을 진행하기 때문에 1년간 52번의 리밸런싱이 일어난다. 52개의 포트폴리오 수익률을 취합하여 연간 수익률로 변환하고, 연간 수익률들을 통해 위의 5가지 주요 지표를 계산한 것은 <Table 1>과 같다. K-평균 군집, 계층적 군집 외에도 벤치마크인 코스피 지수의 매수 후 보유 전략과 공적분을 통한 페어 트레이딩 전략을 같이 비교하였고, 각 지표마다 최고의 값을 진하게 표시하였다.

<Table 1> Annualized profitability metrics

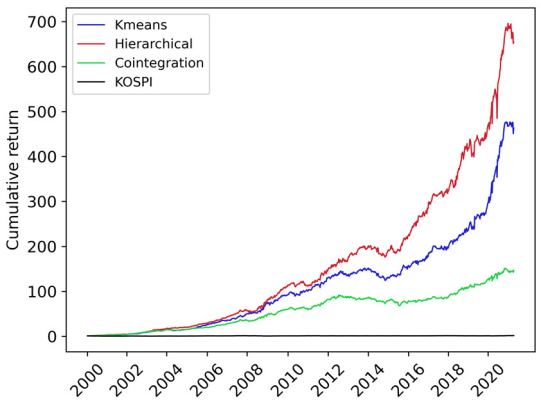
	K-means	Hierarchical	KOSPI	Cointegration
Mean return	**0.305**	0.302	0.054	0.235
Standard deviation	0.141	0.142	0.233	**0.139**
Downside deviation	**0.083**	0.087	0.147	0.086
Sharpe ratio	**2.161**	2.117	0.232	1.692
Sortino ratio	**3.669**	3.476	0.369	2.738
Profit factor	2.109	**2.145**	1.094	1.802
MDD	-0.210	**-0.131**	-0.537	-0.313
Calmar ratio	1.452	**2.303**	0.101	0.753
Gross profit	10.827	10.574	**12.891**	10.004
Gross loss	-5.133	**-4.930**	-11.781	-5.551
Profitable years	**21**	19	17	16
Unprofitable years	**1**	3	5	6

5가지 주요 지표 중에서, 샤프 비율과 소르티노 비율에 대해서는 K-평균 알고리즘이 제일 높았고, 이익 승수, 최대 손실 낙폭, 칼마 비율에 대해서는 계층적 군집 알고리즘이 제일 높은 성과를 보여주었다. 이는 K-평균 군집이 계층적 군집보다 낮은 표준편차를 보여주지만, 시장 상황이 안 좋을 때는 계층적 알고리즘의 수익 방어가 잘 되어, MDD 측면에서 계층적 군집의 성과가 좋기 때문이라고 유추된다. 다른 특이한 점으로는 큰 차이는 아니지만 공적분을 이용한 페어 트레이딩의 표준편차가 가장 작았고, 반면에 코스피 지수의 매수 후 보유 전략은 다른 전략에 비해 아주 높은 표준편차를 기록했다. 코스피 지수는 총 수익 뿐만 아니라 총 손실의 절대값도 가장 커 결과적으로 제일 낮은 이익 승수를 보였다. 수익 기간을 년도로 표현한 지표의 경우, 군집 기반의 페어 트레이딩 전략들이 벤치마크에 비해 더 나은 성과를 보여주었다. 수익성 지표 뿐만 아니라 수익률의 분포도 살펴보았으며 결과는 <Figure 1>과 같다.



<Figure 2> Boxplot of Portfolio Return

코스피 지수의 매수 후 보유 전략은 페어 트레이딩 전략들에 비해서 수익률 분포가 매우 넓고, 평균적으로 낮은 것이 확인된다. 공적분을 통한 페어 트레이딩 전략은 군집 기반 전략들과 비슷한 수익률 분포를 보이지만, 박스가 상대적으로 왼쪽에 위치하여 평균이 낮다. K-평균 군집과 계층적 군집의 평균은 비슷하나, 박스의 길이는 K-평균 군집이 큰 것을 볼 수 있다. 21년간의 데이터를 통해 백테스팅을 수행한 누적 수익률 차트를 그려보았고 최종 결과는 <Figure 2>와 같다.



<Figure 3> Cumulative Return of Backtesting

코스피 지수의 매수 후 보유 전략은 21년동안 69%의 수익률밖에 거두지 못했지만, 페어 트레이딩 전략들은 모두 100배 이상의 수익을 거두었다. 테스트 기간 중 절

반을 지났을 때 공적분 기반 전략은 60배, K-평균은 93배, 계층적 군집은 114배의 수익을 거두었다. 수익의 차이는 더욱 커져 최종 누적 수익률은 각각 147배, 463배, 654배를 기록하였다. 이를 통해, 국내에서 페어 트레이딩 투자전략의 유효성 뿐만 아니라 군집화 기반 페어 탐색의 우월성을 확인할 수 있다.

4.2 기간별 수익성 평가

군집 기반 페어 트레이딩 전략의 누적수익률이 벤치마크보다 아주 높은 숫자를 기록하였지만, 과거에만 높은 수익률을 거두고 비교적 최근에는 유효하지 않을 수 있다. 이에 해당 전략의 수익 지속성을 입증하고자 기간별로 주요 지표들을 살펴보았다. 4년마다의 평균 지표를 K-평균과 계층적 군집에 대해 계산하였고, 결과는 <Table 2>, <Table 3>와 같다.

<Table 2> Sub-period Performance of K-means Clustering

Period	Sharpe ratio	Sortino ratio	Profit factor	MDD	Calmar ratio
2000~2003	4.893	8.870	3.757	-0.164	5.389
2004~2007	2.494	3.930	2.190	-0.210	1.606
2008~2011	1.803	3.173	1.909	-0.101	2.504
2012~2015	0.397	0.634	1.151	-0.181	0.203
2016~2020	1.673	2.660	1.859	-0.083	2.628

K-평균 군집 기반 투자전략의 기간별 주요지표의 경우, 첫 구간인 2000~2003년에서 4가지 지표에 대해 최고의 성과를 나타냈고, 샤프 비율은 4.893을 기록했다. 반면, 네번째 구간인 2012~2015년에는 샤프 비율이 0.397까지 떨어졌고, 최대 손실 낙폭을 제외한 4가지 지표가 모두 가장 낮은 값을 기록했다. 가장 최근 구간인 2016~2020년에는 다행히 이전 구간의 낮은 성과를 회복하였고, 샤프 비율은 <Table 2>에 명시된 KOSPI 수익률보다 꽤 높은 수준인 1.673을 기록했다. 또한, 최대 손실 낙폭은 -8.33%로서 최고의 성과를 보였고, 첫 구간보다는 많이 투자기회가 사라졌지만 여전히 높은 수익성이 있음을 확인하였다.

계층적 군집 기반 투자전략에 대해서도 기간별 주요지표를 계산하였다. K-평균 군집과 비슷하게 첫 구간에서 5가지 지표 모두 최고의 성과를 나타냈고, 샤프 비율은 5.341을 기록했다. K-평균과 마찬가지로 네번째 구간에서 샤프 비율이 0.794까지 떨어져 5가지 지표 모두 최저 성과를 기록하였다. 단, 가장 최근 구간에서 이전 구간의 낮은 성과를 회복하였으며 샤프 비율은 1.430으로 시장 수익률보다는 높은 수준으로 복귀하였다. 즉, 투자기회가 많이 사라졌지만 여전히 높은 수익성이 있음이 확인된다.

<Table 3> Sub-period Performance of Hierarchical Clustering

Period	Sharpe ratio	Sortino ratio	Profit factor	MDD	Calmar ratio
2000~2003	5.341	9.721	4.705	-0.084	11.474
2004~2007	2.583	5.103	2.340	-0.092	3.284
2008~2011	1.550	2.462	1.758	-0.114	1.999
2012~2015	0.794	1.395	1.325	-0.131	0.622
2016~2020	1.430	1.842	1.761	-0.130	1.546

4.3 민감도 분석

군집 알고리즘의 경우 하이퍼파라미터의 값에 따라 각 군집의 결과가 달라질 수 있다. 하이퍼파라미터 최적화를 통해 군집 내 응집력과 군집 간 분리도가 높아지는 방향으로 군집이 선택되지만, 하이퍼파라미터의 값에 따라 결과가 강건하지 않다면 투자전략에 위험이 존재하게 된다. 따라서, K-평균 군집과 계층적 군집에 대한 민감도 분석을 수행하였고, 결과는 <Table 4>, <Table 5>와 같다.

<Table 4> Sensitivity Analysis of K-means Clustering

K	Sharpe ratio	Sortino ratio	Profit factor	MDD	Calmar ratio
5	1.552	2.497	1.713	-0.316	0.921
10	1.975	3.297	1.966	-0.296	1.129
25	2.148	3.650	2.112	-0.270	1.228
50	2.100	3.658	2.115	-0.253	1.294
100	2.161	3.671	2.172	-0.194	1.644
200	2.105	3.478	2.103	-0.211	1.391
400	2.161	3.669	2.109	-0.210	1.452

K-평균 군집 알고리즘의 하이퍼파라미터는 군집의 수이다. 군집의 수를 5, 10, 25, 50, 100, 200, 400으로 설정하여 각각 페어를 탐색하였고, 포트폴리오를 구성하여 수익성 지표를 계산하였다. 군집의 수가 10인 경우 샤프 비율이 2 이하를 기록하였지만, 25 이상의 경우 샤프 비율이 2 이상으로 수렴된 결과를 보였다. 다른 지표들에 대해서도 군집의 수가 25 이상이면 강건한 결과가 나타났고, 군집의 수를 100으로 설정하였을 때 5가지 주요 지표들이 모두 최고의 성과를 기록하였다. 이 때, 이상치를 제거함에 따라 최종적으로 페어 트레이딩에 사용된 군집은 91개였고, 총 760개의 종목 중 614개의 종목이 이상치가 아닌 군집으로 구성되었다. 각 군집에서 스프레드가 포착된 종목은 201개의 종목이었고, 이는 총 종목 수 대비 26%에 해당하는 종목들이 투자 전략에 사용

되었음을 의미한다. 가장 큰 군집에 속한 종목 수는 20개였고, 91개의 군집에 종목들이 고루 퍼져있는 것을 확인할 수 있었다.

<Table 5> Sensitivity Analysis of Hierarchical Clustering

	Sharpe ratio	Sortino ratio	Profit factor	MDD	Calmar ratio
0.1	1.886	3.054	1.916	-0.229	1.279
0.2	2.029	3.394	2.041	-0.186	1.560
0.3	2.094	3.461	2.124	-0.156	1.902
0.4	2.117	3.476	2.145	-0.131	2.303
0.5	1.969	3.184	2.048	-0.147	1.935
0.6	1.914	3.108	2.005	-0.157	1.766
0.7	1.931	3.063	2.030	-0.176	1.605
0.8	1.897	2.990	2.012	-0.195	1.433
0.9	1.881	2.970	1.996	-0.229	1.223

계층적 군집 알고리즘의 하이퍼파라미터는 이상치 임계값의 백분위수 α이다. α를 0.1에서 0.9까지 0.1 단위로 나눠 각각 페어를 탐색 후 포트폴리오를 구성하여 수익성 지표를 계산하였다. α가 0.2, 0.3, 0.4의 경우 샤프 비율이 2 이상을 기록하였고, α가 0.4일 때 5가지 지표들이 모두 최고의 성과를 기록하였다. 이 때, 이상치를 제거함에 따라 최종적으로 페어 트레이딩에 사용된 군집은 169개였고, 총 760개 종목 중 677개의 종목이 이상치가 아닌 군집으로 구성되었다. 각 군집에서 스프레드가 포착된 종목은 199개였고, K-평균 군집 알고리즘의 경우와 비슷한 양의 종목들이 투자 전략에 사용되었다. 가장 큰 군집에 속한 종목 수는 94개였고, 상위 군집에 많은 종목들이 밀집되어 있는 것을 확인할 수 있었다. α의 값이 0.3과 0.4의 경우 지표를 비교해보면 강건한 결과를 보여주지만, 0.4와 0.5의 경우에는 상대적으로 민감한 변화가 나타났다. 모두 높은 수익성을 보여주지만, 계층적 군집의 경우 이러한 민감도를 고려한 하이퍼파라미터 최적화가 필요하다.

5. 결 론

페어 트레이딩에서 가장 중요한 단계는 페어를 포착하는 단계이고, 해당 연구에서는 군집화를 통해 유사한 가격흐름의 페어를 탐색하였다. 스프레드가 커지는 페어에 대해 포트폴리오에 편입하였고, 해당 전략이 국내 시장에서 유효한지 확인하고자 2000년부터 2020년까지의 코스피 지수에 속한 주가 데이터로 백테스팅을 진행하였다. 백테스팅 결과 K-평균 군집의 경우 2.161의 높은 샤프 비율을 기록하였고, 계층적 군집은 21년간 654배의 누적 수익률을 기록하였다. 기간별로 나누어 수익성을 살펴보아도 최근 구간인 2016년에서 2020년 동안 각각 1.673과 1.430의 높은 샤프 비율을 기록하여 여전히 국내 시장에서 페어 트레이딩의 투자기회가 남아 있음을 확인하였다. 또한, 코스피 지수의 총 종목 수 대비 26%에 해당하는 종목만이 해당 전략에 사용되었으므로 투자규모가 늘어나더라도 적용가능한 투자전략임을 보여준다. 군집의 하이퍼파라미터에 대한 민감도 분석에도 강건한 수익성을 보여주어 해당 전략의 유효성을 입증하였다. 하지만, 모든 종목에 대해 공매도를 할 수 있는 것은 아니고, 1주일마다 포트폴리오의 리밸런싱을 한다면 공매도의 거래비용도 상당할 것으로 예상되기에, 추후 연구로는 공매도의 제한과 거래비용 등을 고려한 투자기회에 관하여 고찰이 필요할 것이다.

경기 침체가 우려되고 자산변동성이 확대되는 상황에서 통계적 차익거래는 시장 중립적 투자전략으로 자산의 변동성에 덜 영향을 받는다. 연구결과의 상용화를 통하여 핀테크 영역에서의 활용이 유의미할 것으로 사료된다. 또한 금융 시계열 데이터에 적합한 군집화 알고리즘을 통해 다른 투자전략으로의 응용 가능성도 제시된다. 추후 연구로는 군집화 지표에서 수익성 지표로 인과관계가 있는지 그레인저 인과관계를 통해 확인할 수 있다. 랜드 비율[20], 상호정보량[25]과 같은 군집화 지표를 추가적으로 측정할 수 있고, 군집화 지표와 수익성 지표가 모두 최대화되도록 지속적인 학습이 가능한 모델을 구성할 수 있다. 시계열 데이터에 특화된 군집화 알고리즘의 적용이 가능하고, 동적 시간 워핑 등의 데이터 방식[18], 설명가능한 섀플리 값 등의 특징 방식[21], 자기조직화지도 등의 딥러닝 방식[16]을 포함한다[2].

Acknowledgement

This work has been supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1F1A1062917).

References

[1] Aggarwal, C.C., Hinneburg, A., and Keim, D.A., On the surprising behavior of distance metrics in high dimensional space, In *International Conference on Database Theory*, 2001, Springer, Berlin, Heidelberg, pp. 420-434.

[2] Alqahtani, A., Ali, M., Xie, X., and Jones, M.W., Deep Time-Series Clustering: A Review, *Electronics*, 2021, Vol. 10, No. 23, pp. 3001.

[3] Aranganayagi, S. and Thangavel, K., Clustering categorical

data using silhouette coefficient as a relocating measure, In *International conference on computational intelligence and multimedia applications*, 2007, IEEE ,Vol. 2, pp. 13-17.

[4] Arthur, D. and Vassilvitskii, S., *k-means++: The advantages of careful seeding*, Stanford, 2006.

[5] Caldeira, J. and Moura, G.V., Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy, Available at SSRN 2196391, 2013.

[6] Day, W.H. and Edelsbrunner, H., Efficient algorithms for agglomerative hierarchical clustering methods, *Journal of Classification*, 1984, Vol. 1, No. 1, pp. 7-24.

[7] Dickey, D.A. and Fuller, W.A., Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association*, 1979, Vol. 74, No. 366a, pp. 427-431.

[8] Flori, A. and Regoli, D., Revealing pairs-trading opportunities with long short-term memory networks, *European Journal of Operational Research*, 2021, Vol. 295, No. 2, pp. 772-791.

[9] Goldkamp, J. and Dehghanimohammadabadi, M., Evolutionary multi-objective optimization for multivariate pairs trading, *Expert Systems with Applications*, 2019, Vol. 135, pp. 113-128.

[10] Gupta, K. and Chatterjee, N., Selecting stock pairs for pairs trading while incorporating lead–lag relationship, *Physica A: Statistical Mechanics and its Applications*, 2020, Vol. 551, 124103.

[11] Han, C., He, Z., and Toh, A.J.W., Pairs Trading via Unsupervised Learning, Available at SSRN 3835692, 2021.

[12] Hartigan, J.A. and Wong, M.A., Algorithm AS 136: A k-means clustering algorithm, *Journal of the royal statistical society. series c (applied statistics)*, 1979, Vol. 28, No. 1, pp. 100-108.

[13] Hautamäki, V., Cherednichenko, S., Kärkkäinen, I., Kinnunen, T., and Fränti, P., *Improving k-means by outlier removal, In Scandinavian conference on image analysis*, Springer, Berlin, Heidelberg, 2005, pp. 978-987.

[14] Johnson, S.C., Hierarchical clustering schemes, Psychometrika, 1967, Vol. 32, No. 3, pp. 241-254.

[15] Magdon-Ismail, M. and Atiya, A.F., Maximum drawdown, *Risk Magazine*, 2004, Vol. 17, No. 10, pp. 99-102.

[16] Manduchi, L., Hüser, M., Vogt, J., Rätsch, G., and Fortuin, V., DPSOM: Deep Probabilistic Clustering with self-organizing maps, arXiv preprint arXiv:1910.01590, 2019.

[17] Murtagh, F., A survey of recent advances in hierarchical clustering algorithms, *The Computer Journal*, 1983, Vol. 26, No. 4, pp. 354-359.

[18] Müller, M., Dynamic time warping, *Information Retrieval for Music and Motion*, 2007, pp. 69-84.

[19] Ramos-Requena, J.P., Trinidad-Segovia, J.E., and Sánchez-Granero, M.A., Introducing Hurst exponent in pair trading, *Physica A: Statistical Mechanics and its Applications*, 2017, Vol. 488, pp. 39-45.

[20] Rand, W.M., Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*, 1971, Vol. 66, No. 336, pp. 846-850.

[21] Roth, A.E., *The Shapley Value: Essays in Honor of Lloyd S*, Shapley, Cambridge University Press, 1988.

[22] Sarmento, S.M. and Horta, N., Enhancing a pairs trading strategy with the application of machine learning, *Expert Systems with Applications*, 2020, Vol. 158, 113490.

[23] Sharpe, W.F., The Sharpe ratio, *Streetwise–the Best of the Journal of Portfolio Management*, 1998, pp. 169-185.

[24] Wen, D., Ma, C., Wang, G.J., and Wang, S., Investigating the Features of Pairs Trading Strategy: A Network Perspective on the Chinese stock market, *Physica A: Statistical Mechanics and its Applications*, 2018, Vol. 505, pp. 903-918.

[25] Zhang, H., Ho, T.B., Zhang, Y., and Lin, M.S., Unsupervised feature extraction for time series clustering using orthogonal wavelet transform, *Informatica*, 2006, Vol. 30, No. 3.

ORCID

Poongjin Cho | https://orcid.org/0000-0002-1844-0472
Minhyuk Lee | https://orcid.org/0000-0002-1838-5821
Jae Wook Song | https://orcid.org/0000-0001-6455-6524