



KNU DART

Dart Quant Strategy Series

Many-to-Many Pairs Trading Strategy

Choi Munseok, 12th
Choi Wonjun, 13th

**May
2024**

목차

I. 소개	1
II. 방법론과 데이터	2
2.1 주성분분석	3
2.2 군집화 알고리즘	4
2.3 성능 평가	5
2.4 칼만필터(Kalman Filter)	6
2.5 정상성 검정	7
2.6 다중 페어트레이딩	8
2.7 데이터 설명	9
III. 실증연구	10
3.1 분석기간	11
3.2 군집화 기반 양상블 페어선택 알고리즘	12
3.3 칼만필터를 활용한 스프레드 동적 추정 알고리즘	13
IV. 결론	14
V. 부록	15

DART 2024 First Semester Pairs Trading Project

Choi Munseok
cmschs0301@knu.ac.kr

Choi Wonjun
song1002@knu.ac.kr

[Github]

<https://github.com/DART-KNU/Pairs-Trading-from-Beginning-to-End>

I. Introduction

페어 트레이딩(Pairs Trading)은 금융 시장에서 오랜 역사를 가진 시장 중립적 투자 전략이다. 이 전략은 두 개의 유가증권이 서로의 가격 변동을 따르는 경향이 있다는 가정하에, 그들 간의 가격 차이를 이용하여 수익을 창출한다. 한 증권이 과대평가되면 매도하고, 다른 증권이 과소평가되면 매수하여 가격 차이가 정상으로 돌아올 때 이익을 실현한다. 이러한 전략은 시장의 방향성과 무관하게 안정적인 수익을 추구할 수 있다. 본 연구는 페어 트레이딩 전략의 전 과정을 다루며, 다양한 페어 선택 및 거래 전략에서 기존 연구들의 한계점을 개선하여 효율성을 극대화하고자 한다. 전통적인 페어 선택 방법은 모든 가능한 유가증권 쌍을 고려해야 하므로 계산 비용이 높고, 다중 검정 문제로 인해 신뢰성 있는 페어를 선택하기 어려운 단점이 있다. 따라서 본 연구에서는 Simão Moraes Sarmento[9]에서 제안된 군집화 기반 앙상블 페어 선택 알고리즘과, 이를 팩터 모델의 관점에서 개선한 알고리즘을 사용하여 다중 검정 문제를 완화하였다. Krauss[3]는 개별 주식의 모멘텀이 강하게 나타날 경우, 두 주식 간의 가격 차이가 커져 스프레드가 확대될 수 있음을 분석하였다. 반면, 지수와 지수는 모멘텀보다 평균 회귀 특성이 강하게 나타난다. 따라서 본 연구에서는 앞서 제안된 군집화 기반 앙상블 페어 선택 알고리즘을 통해 만들어진 군집 내에서 다대다 페어 트레이딩(Many-to-Many Pair Trading; MMPT)을 진행한다. Wang[7]에서 제안한 MMPT는 군집 내 일부 one-to-one 페어가 소실되는 문제가 있다. 본 연구에서는 군집 내 one-to-one 페어를 최대한 보존하는 서브군집 분할 알고리즘을 사용하여 이를 해결한다. 또한 군집이 깨지는 경우를 방지하기 위해 MMPT에서 칼만 필터를 활용해 스프레드를 동적으로 추정하였다. 이러한 일련의 전략을 바탕으로 롤링 윈도우 분석(Rolling Window Analysis)을 통해 전략의 성능을 백테스트 하였다.

II. Methodology and Data

2.1 주성분 분석(Principal component analysis)

주성분 분석(Principal Component Analysis; PCA)은 널리 사용되는 차원 축소 및 특징 추출 기법이다. PCA의 주요 목표는 상관된 변수들의 큰 집합을 상관되지 않은 소수의 주성분으로 변환하여, 원본 데이터의 변동성을 최대한 유지하는 것이다. PCA는 데이터가 가장 많이 변동하는 방향(주성분)을 식별하여 이 방향으로 데이터를 투영한다. 이러한 주성분들은 서로 직교하므로, 추출된 특징들이 상관되지 않도록 보장한다. PCA는 공분산 행렬의 고유값 분해 또는 데이터 행렬의 특이값 분해(Singular Value Decomposition, SVD)에 기반한다. 본 연구에서는 수치적 안정성과 효율성 때문에 SVD 접근법을 사용한다.

2.1.1 특이값 분해(SVD)

주어진 데이터 행렬 (A)가 ($m \times n$) 크기일 때, 여기서 (m)은 관측치의 수, (n)은 변수의 수를 나타낸다. SVD는 (A)를 다음과 같이 세 개의 행렬로 분해한다:

$$A = U\Sigma V^T$$

- U 는 ($m \times m$) 크기의 직교 행렬로, 왼쪽 특이벡터를 포함한다.
- Σ 는 ($m \times n$) 크기의 대각 행렬로, 준양정부호 실수인 특이값들을 포함한다.
- V 는 ($n \times n$) 크기의 직교 행렬로, 오른쪽 특이벡터를 포함한다.

2.1.2 SVD를 활용한 PCA

1) 표준화: 데이터를 평균이 0이고 분산이 1이 되도록 정규화 한다. 이는 PCA가 변수의 척도에 민감하기 때문에 중요하다.

$$X_{\text{normalized}} = \frac{X - \mu}{\sigma}$$

여기서 (X)는 원본 데이터 행렬, (μ)는 평균, (σ)는 표준편차이다.

2) 공분산 행렬 계산: 정규화 된 데이터 행렬의 공분산 행렬 (C)를 계산한다.

$$C = \frac{1}{m-1} X_{\text{normalized}}^T X_{\text{normalized}}$$

3) 특이값 분해: 공분산 행렬 (C)에 대해 SVD를 수행한다.:

$$C = U \Sigma V^T$$

4) 주성분 도출: 주성분은 공분산 행렬의 고유벡터로, 이는 (V)의 열들이다. 해당 고유값(대각 성분(Σ))은 각 주성분이 설명하는 분산을 나타낸다.

5) 주성분으로 투영: 원본 데이터를 주성분 공간으로 변환이다.

$$Z = X_{\text{normalized}} V$$

여기서 Z 는 주성분 점수 행렬이다.

2.2 군집화 알고리즘

2.2.1 K-Means Clustering

K-Means 군집화는 데이터를 여러 개의 군집으로 나누는 비지도 학습 알고리즘이다. 각 군집은 군집의 중심(centroid)에서 가장 가까운 데이터 포인트들로 구성되며, 이 중심은 각 군집 내의 데이터 포인트들의 평균 위치로 계산된다. K-Means 알고리즘의 기본 단계는 다음과 같다:

- 1) 초기화: 엘보우(Elbow) 방법을 사용하여 군집의 수(K)를 설정한다. 이 방법은 군집 내 거리 제곱합(Sum of Squared Distances, SSD)을 설정된 k 값에 따라 계산하고, 그래프를 통해 가장 큰 기울기 변화가 발생하는 지점을 찾아 최적 군집 k 값으로 설정한다.
- 2) 할당: 각 데이터 포인트를 가장 가까운 중심에 할당한다. 이는 유클리드 거리와 같은 거리 측정 방법을 사용하여 수행되어진다.
- 3) 중심 업데이트: 각 군집의 새로운 중심을 계산한다. 이는 군집 내 모든 데이터 포인트의 평균 위치로 재설정 되어진다.
- 4) 수렴 조건 검사: 중심의 위치가 더 이상 변화하지 않거나 설정된 반복 횟수에 도달할 때까지 2)와 3)단계를 반복한다.

2.2.2 Hierarchical Clustering

계층적 군집화는 데이터 포인트들을 계층적으로 병합하여 군집을 형성하는 비지도 학습 방법 중 하나이다. 이 연구에서는 주성분 분석(PCA)으로 차원 축소된 데이터를 군집화하는 데 계층적 군집화 방법을 사용하였다. 이 방법은 데이터 포인트들을 개별 군집으로 시작하여 점차적으로 가장 가까운 군집들을 병합해 나가는 하향식 접근 방식을 따른다. 먼저, 이 방법에서는 원하는 군집 수를 지정한 후, 주어진 데이터에 대해 모델을 피팅하고 알고리즘을 적용한다. 군집 수를 결정하는 것은 계층적 군집화에서 중요한 단계이다. 본 연구에서는 데이터 포인트 수의 제곱근 값을 최적의 군집 수로 설정하였다.

2.2.3 DBSCAN

DBSCAN(Density based spatial clustering of applications with noise)의 경우 앞서 설명한 K-means와 Hierarchical 방식과 다르게 군집간의 거리가 아닌 밀도를 기반으로 클러스터링을 진행하는 방식이다. 쉽게 말해 한 점을 기준으로 반경 ϵ 내에 점 n 개 이상이 있으면 하나의 군집으로 인식하는 방식이다. 이는 계층적 군집화 방식과 유사하게 미리 클러스터 개수를 설정할 필요가 없으며, 이상치 탐지가 쉽다는 장점이 있다. 그러나 고차원 데이터 혹은 Sparse한 데이터에서는 성능이 잘 나오지 않는다는 단점이 존재한다. DBSCAN의 기본 단계는 다음과 같다.

- 1) 매개변수 설정: ϵ (이웃으로 설정할 최대 반경)와 minPts(군집의 최소 포인트가 되기 위한 최소 이웃 점의 수)를 설정한다.
- 2) 이웃 탐색: 각 데이터 포인트에 대해 반경 ϵ 이내이면서 이웃의 수가 minPts 이상인 핵심 포인트를 찾는다.
- 3) 군집 형성: 핵심 포인트를 중심으로 이웃 포인트를 재귀적으로 검토하며 군집을 확장한다.
- 4) 노이즈 점 식별: 어떤 군집에도 속하지 않는 점은 노이즈로 분류한다.
- 5) 반복: 모든 점이 군집에 속하거나 노이즈로 분류될 때까지 반복한다.

2.3 성능 평가

2.3.1 결정계수(R-Squared)를 활용한 군집화 성능 평가

결정계수(R^2)는 회귀 분석과 모델 성능 평가에서 널리 사용되는 통계적 지표이다. 이는 상관계수의 제곱으로 계산되며, 독립 변수들이 종속 변수의 변동성을 얼마나 잘 설명하는지를 나타낸다.

R^2 는 0에서 1 사이의 값을 가지며, 값이 1에 가까울수록 모델이 데이터의 변동성을 잘 설명하고 있음을 의미한다. 이는 두 변수 간의 강한 선형 관계를 나타낸다. 반면, R^2 값이 0에 가까울수록 설명력이 낮아 모델의 예측력이 부족함을 나타낸다. 결정계수는 다양한 데이터 분석 및 모델 평가 상황에서 중요한 역할을 한다. 페어 트레이딩에서 군집화의 성능을 평가하는 데 결정계수(R^2)를 사용하는 것은 각 군집 내에서 선택된 페어가 얼마나 일관된 관계를 가지는지를 평가하는 데 매우 유용하다. 군집화 알고리즘을 통해 비슷한 특성을 가진 자산들을 동일한 군집으로 묶은 후, 해당 군집 내의 자산 쌍들 간의 결정계수를 계산하여 군집의 품질을 평가한다. 높은 R^2 값은 해당 군집 내 자산들이 강한 선형 관계를 가지고 있음을 의미하며, 이는 페어 트레이딩 전략의 성공 가능성을 높인다. 따라서, 군집화된 자산들 간의 결정계수를 통해 군집화의 효율성을 평가하고, 최적의 페어를 선택하는 데 중요한 기준으로 활용할 수 있다. 이러한 접근 방식은 페어 트레이딩 전략의 신뢰성을 높이고, 시장 변동성에 대한 민감도를 줄이는 데 기여한다.

2.4 칼만필터(Kalman Filter)

칼만 필터(Kalman Filter)는 루돌프 E. 칼만에 의해 1960 년에 제안된 알고리즘으로, 바로 이전 시간에 추정된 값을 토대로 해서 현재의 값을 추정하는 재귀필터이다. 값을 예측할 뿐만 아니라 오차 공분산을 함께 예측하여 노이즈를 잘 반영한다는 특징을 가지고 있다. 따라서 노이즈가 많이 포함된 주가와 같은 모델에 특히 효과적이라 볼 수 있다.

칼만필터의 단계는 예측단계와 보정단계를 거쳐 최종적으로 다음 상태를 예측한다. 예측단계에서는 현재 상태와 제어 입력 값을 사용하며 다음 시각의 상태와 오차공분산의 값을 추정한다. 그 후 보정단계에서는 측정값과 예측값의 차이를 보정해서 새로운 추정값을 계속 업데이트 해 나간다.

예측 단계 (Prediction Step)

상태 예측: 이전 상태와 현재 제어 입력을 사용하여 시스템의 다음 상태를 예측한다. 이는 상태 전이 모델을 사용하여 수행되며 이후 보정 단계를 거쳐 예측값은 보정된다.

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_k \hat{\mathbf{x}}_{k-1|k-1} + \mathbf{B}_k \mathbf{u}_k$$

오차 공분산 예측: 상태 예측과 함께, 예측된 상태에 대한 불확실성(오차)을 계산한다. 이는 예측된 상태가 얼마나 신뢰할 수 있는지를 나타내며, 오차 공분산이 클수록 추정 오차가 크고 작을 수록 오차가 작다.

$$\mathbf{P}_{k|k-1} = \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}_k^T + \mathbf{Q}_{k-1}$$

우리의 모델에서는 제어 입력이 존재하지 않으므로 B_k 를 0 으로 할당하였으며 상태 전이행렬 F_k 의 경우 가격만을 가지고 다음 상태를 예측하도록 설정하였다.

보정 단계 (Update Step)

보정 단계에서는 새로운 측정값을 이용하여 예측된 상태를 보정한다. 보정 단계는 세 부분으로 나눌 수 있다.

칼만이득 계산: 칼만 이득은 예측된 상태와 실제 측정값을 얼마나 반영할지를 결정하는 가중치이다. 이는 예측된 상태의 불확실성과 측정값의 불확실성을 비교하여 계산된다. 칼만이득은 예측값을 얼마나 보정할지 되는 인자가 된다.

$$\mathbf{S}_k = \mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^T + \mathbf{R}_k$$

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{H}_k^T \mathbf{S}_k^{-1}$$

칼만 이득 공식을 보면 알 수 있듯이 R_k 값이 커지게 된다면 칼만 이득은 작아지게 된다. 또한 앞서 오차 공분산 예측에서 P_k 에 노이즈 Q 가 들어가므로 Q 값이 커지게 된다면 칼만이득의 값도 또한 작아진다.

상태 갱신: 예측치를 통해 계산된 예측값 과 실제 측정값의 차이를 바탕으로 예측값을 보정해 최종 추정값을 계산한다.

$$\tilde{\mathbf{y}}_k = \mathbf{z}_k - \mathbf{H}_k \hat{\mathbf{x}}_{k|k-1}$$

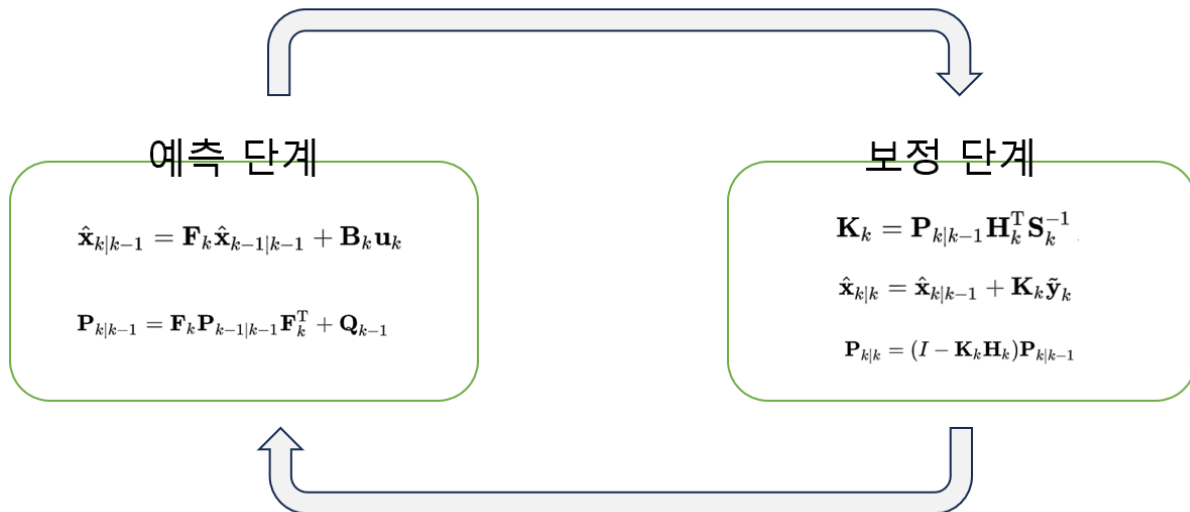
$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \tilde{\mathbf{y}}_k$$

추정된 계산식을 보면 칼만 이득이 커지면 추정 값 계산에 예측값이 반영되는 비율이 커지고 작아지면 측정값이 반영되는 비율이 커지게 됨을 알 수 있다.

오차 공분산 갱신: 상태 갱신 후, 새로운 상태에 대한 불확실성을 갱신한다. 이는 갱신된 상태가 얼마나 신뢰할 수 있는지 판단하는 척도를 나타낸다.

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_{k|k-1}$$

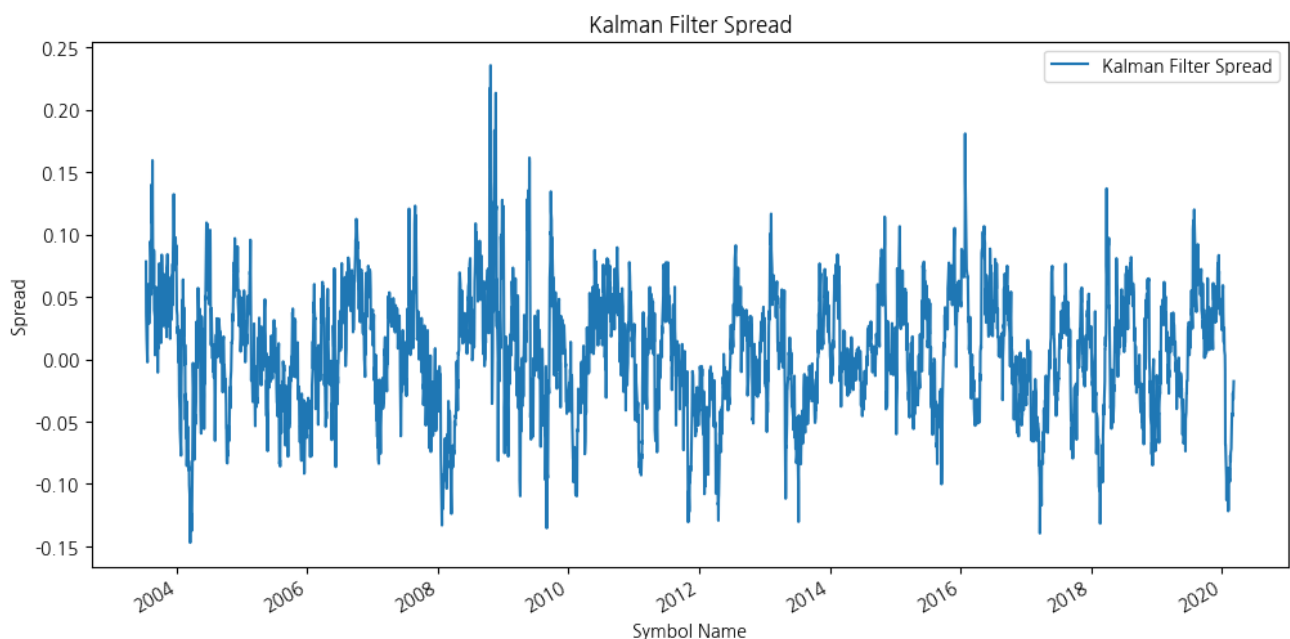
이와 같이 칼만필터는 예측단계와 보정 단계를 재귀적으로 반복해 연속적으로 데이터를 분석하며 시스템의 상태를 점점 더 정확하게 추정할 수 있다. 이 필터는 특히 예측 정확도가 중요한 금융 및 엔지니어링 분야에서 중요한 도구로 활용되곤 한다. 전체 수식의 구조도는 <그림 1>과 같다.



<그림1 > 칼만필터의 구조도

2.5 정상성 검정

시계열에서의 정상성은 일반적으로 시간이 지나도 평균, 분산 및 자기 상관 구조가 변하지 않는것을 의미한다. 즉 시계열에 추세, 주기 또는 시간에 따라 변하는 체계적인 패턴이 없음을 의미한다. 페어트레이딩의 경우 두 가격 사이의 잔차가 정상성을 가지고 있어야 한다. 그렇지 않다면 두 페어 사이의 공적분 관계가 없다는 것을 의미하며, 두 시계열 변수가 장기적으로 관련되어 있지 않다. 라고 간주한다. 정상성이 있는 그래프의 예시는 <그림2>와 같다.



<그림2> 정상시계열 예시 데이터

따라서 스프레드가 정상성 기질이 있는지 검사 하는것이 중요하다. 이는 가격 시계열이 일시적으로 평균에서 벗어난 후에 평균값으로 다시 돌아가려는 경향이 존재함을 의미하며 가격 시계열이 일정한 수준을 중심으로 진동하며 이 수준에서의 이탈은 일시적이며 다시 평균으로 돌아간다는 뜻을 내포하고 있다. Pd.Chan[4]의 저서에 따르면 정상성의 검정은 크게 ADF 검정을 통해 확인할 수 있다.

2.5.1 ADF 검정

ADF 검정의 식은 다음과 같다.

$$\Delta Y(t) = \lambda Y(t-1) + \mu + \beta t + \alpha_1 \Delta Y(t-1) + \dots + \alpha_k \Delta Y(t-k) + \varepsilon_t$$

이때 주식시장의 일간 변동성은 상당히 높으며 추세에 비해 훨씬 큰 영향을 미치므로 β 를 0로 가정하여 모형을 간단하게 유지하고 변동성에 집중하게 된다.

$$\Delta Y(t) = \lambda Y(t-1) + \mu + \varepsilon_t$$

ADF검정은 시계열이 단위근 λ (시계열 데이터가 시간 경과에 따라 그 자체의 과거 값에 의존하는 정도가 1인 경우)가 1인지 판단하는데 사용되어진다. 만약 단위근이 존재한다면 정상성을 가지지 않으며, 스프레드가 발산하거나 수렴할 경우가 있음을 나타낸다. 단위근의 값에 따라 우리는 스프레드가 0으로 수렴하고 있는지 확인할 수 있다.

회귀계수 $\lambda = 0$ 이 아니라면 $\Delta y(t)$ 가 $y(t-1)$ 에 영향을 받으며, $\lambda > 0$ 이면 추세가 존재하여 주가가 평균회귀하지 않고, $\lambda < 0$ 이면 주가가 평균회귀할 가능성이 있다. 이때 회귀 계수 λ 를 표준오차 $SE(\lambda)$ 로 나눈 $\lambda/SE(\lambda)$ 를 검정통계량으로 사용하여 임계값보다 낮으면 주가가 정상성을 띄며 평균으로 회귀할 것으로 본다.

2.6 MMPT(Many-to-Many Pair Trading)

Wang[7]에서 제안한 MMPT(Many-to-Many Pair Trading)는 일대일 페어 트레이딩을 다대다 페어 트레이딩으로 확장하는 프레임워크이다. 이는 연관 규칙 알고리즘, OPTICS 클러스터링 알고리즘, 그리고 이분 그래프 분할 알고리즘을 사용하여 다대다 페어를 형성한다. 그러나 Wang[7]의 MMPT는 연관 규칙 알고리즘과 이분 그래프 분할 알고리즘을 사용함에 따라 군집 내 일부 일대일 페어가 소실되는 문제가 있다. 따라서 본 연구에서는 군집 내 일대일 페어를 최대한 보존하는 서브군집 분할 알고리즘을 사용하여 이를 해결하였다. 또한, 칼만 필터를 활용한 스프레드 동적 추정 알고리즘과 Many-to-Many 페어 스프레드 추정을 통해 더욱 효율적인 MMPT를 구현하였다. 이러한 방법론을 통해 기존의 한계를 극복하고 페어 트레이딩 전략의 효율성을 향상시키고자 하였다.

2.7 데이터 설명

FNGuide에서 제공하는 1999.05.14부터 2024.05.13까지의 KOSPI 종목 수정주가 가격 데이터를 사용하였다. 기간내 신규 상장/상장 폐지된 종목들은 제외하고 분석을 진행하였다.

III. 실증연구

3.1 분석기간

1999.05.14부터 2024.05.13까지의 KOSPI 종목 수정주가 가격 데이터를 Rolling Window방식으로 실증연구를 진행한다. 각 Window 마다 Train/Test 데이터셋으로 8:2 비율로 나눠 연구를 진행하였다. '5.3 전략 백테스트'에서와 같이 5개의 Window로 분할하였다.

	Window1	Window2	Window3	Window4	Window5
Train 시작날짜	1999.05.14	2000.05.25	2001.06.14	2002.07.02	2003.07.14
Train 종료날짜	2016.01.04	2017.01.16	2018.02.02	2019.02.25	2020.03.09
Test 시작날짜	2016.01.05	2017.01.17	2018.02.05	2019.02.26	2020.03.10
Test 종료날짜	2020.03.17	2021.03.31	2022.04.13	2023.04.24	2024.05.13

<표 1> 각 윈도우에 설정된 날짜 분포

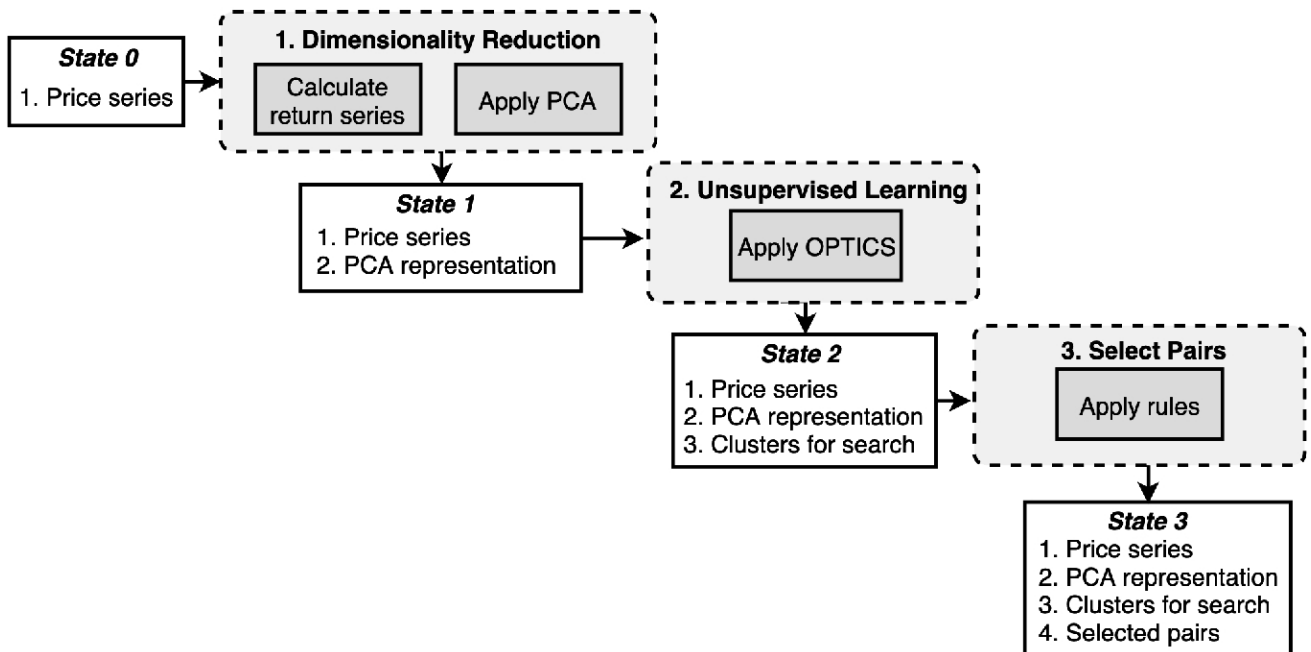
3.2 군집화 기반 앙상블 페어선택 알고리즘

일반적으로 페어트레이딩에서 페어선택 방법은 모든 유가증권 간의 조합을 고려하여 모든 가능한 후보 페어를 선택하는 것이다. 사용가능한 유가증권이 n 개라고 했을 때, 가능한 조합의 수는 $\frac{n*(n-1)}{2}$ 이다. 이때, 두가지 문제가 발생하게 된다. 첫째, 모든 가능한 조합에 대한 평균회귀 테스트의 시간복잡도는 $O(n^2)$ 이다. 따라서 고려되는 유가증권이 늘어날수록, 계산량이 급격하게 증가하게 된다. 둘째, 다중 검정 문제가 발생한다. 하나의 테스트에서 귀무가설을 기각할 확률, 즉 제 1종 오류를 저지를 확률은 α 이다. 하나의 테스트에서 귀무가설을 기각하지 않을 확률은 $1 - \alpha$ 이며, 모든 m 개의 테스트에서 귀무가설을 기각하지 않을 확률은 $(1 - \alpha)^m$ 이다. 적어도 하나의 테스트에서 귀무가설을 기각할 확률은 $1 - (1 - \alpha)^m$ 가 된다. 즉, 각각의 가설에 대해 유의수준 α (1종오류)인 검정을 동시에 수행할 경우 전체 오류율이 매우 커지게 된다. 해당부분은 Family-wise type 1 error rate(FWER)로 잘 알려져 있다. 다중 비교 문제의 영향을 완화하는 방법은 두가지이다. 첫째, 본페로니 교정(Bonferroni correction)과 같은 다중 보정테스트를 적용하는

방법이다. 본페로니 교정은 유의수준을 총 검정의 수로 나누어서 개별 검정을 보정하는 방법이다. 하지만, Harlacher[8]는 본페로니 보정이 페어선택에 있어 너무 보수적이 되어 실제로 페어의 발견을 방해한다는 것을 밝혔다. 둘째, 통계적 테스트의 수를 줄이는 것이다. 본 장에서는 Simão Moraes Sarmento[9]가 제안한 군집화 기반 앙상블 페어선택 알고리즘을 사용하여 통계적 테스트의 수를 줄인다. 또한, 해당 알고리즘에 대한 저자의 개인적인 생각을 바탕으로 수정한 새로운 방법론과 성능을 비교한다. 군집화 기반 앙상블 페어선택 알고리즘은 (1) 비슷한 위험요소에 노출된 유가증권끼리 묶여 함께 움직일 가능성이 높으며 (2) 필요한 통계적 테스트의 수가 급감하는 장점이 있다.

3.2.1 알고리즘1

Simão Moraes Sarmento[9]가 제안한 알고리즘은 <그림3>과 같이 비지도 학습 알고리즘을 사용하여 의미 있는 자산 군집을 추론하여 쌍을 선택하는 것이다. 순서는 다음과 같다.



<그림3> Simão Moraes Sarmento[9]가 제안한 페어 선택 프레임워크

1) 수익률 시계열 계산: 증권 (i)의 시점 (t)에서의 수익률 시계열은 가격 시계열에서 도출된다.

$$r_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}$$

여기서($P_{i,t}$)는 시점(t)에서의 증권(i)의 가격이다.

2) 정규화: 수익률 시계열은 평균(μ_i)를 빼고 표준편차(σ_i)로 나누어 정규화 한다:

$$z_{i,t} = \frac{r_{i,t} - \mu_i}{\sigma_i}$$

3) 상관 행렬 계산: 모든 자산의 정규화된 수익률 시계열을 사용하여 상관 행렬 (C)를 계산한다:

$$C_{i,j} = \frac{1}{T-1} \sum_{t=1}^T z_{i,t} z_{j,t}$$

4) 고유벡터 및 고유값: 상관행렬(C)의 고유벡터와 고유값을 특이값분해(SVD)를 사용하여 구한다. 모든(n)개의 증권에 대한 정규화된 수익률 시계열을 행렬(A)에 쌓아 다음과같이 분해한다:

$$A = USV^T$$

여기서 (U)는 왼쪽 특이벡터를 포함하는 직교 행렬, (S)는 내림차순으로 정렬된 특이값을 포함하는 대각 행렬, (V)는 오른쪽 특이벡터를 포함하는 전치 직교 행렬이다.

5) 특징 벡터 선택: 가장 큰 고유값(최대 변동성)에 해당하는 고유벡터를 선택한다. 필요한 분산 설명 수준에 따라 고유벡터의 수 (k)를 선택한다.

6) 새로운 데이터셋 형성: 원래행렬(A)를 특징벡터로 곱하여 차원이 축소된 새로운 데이터셋을 생성한다:

$$A_{new} = A \cdot \text{Feature Vector}$$

결과 행렬의 크기는 (n × k)이며, (k)는 선택된 주성분의 수이다.

7) 주성분의 수(k)는 일반적으로 각 성분이 설명하는 총 분산의 비율을 분석하고, 원하는 분산 비율을 설명하는 성분의 수를 선택하여 결정한다. 이 맥락에서 차원의 저주와 높은 차원에서의 클러스터링 비효율성 문제를 고려해야 한다. 카이저-가트만 규칙(Kaiser-Guttman Criterion)의 고유값 1 기준에 따라 고유값이 1 이상인 고유벡터만 선택한다. 또한 Berkhin [10]에 따르면 PCA 차원의 수는 15로 제한된다.

8) OPTICS를 적용해 군집을 형성하고, 군집 내 페어를 선정한다. 다만 본 연구에서는 OPTICS 대신 DBSCAN, K-Means, Hierarchical 군집화 방법론을 사용한다.

3.2.2 알고리즘2

본 연구에서 제안하는 알고리즘은 팩터 모델의 관점으로 특잇값 분해(Singular Value Decomposition, SVD)를 진행하고, 팩터모델의 민감도 행렬(n × k)를 사용하여 군집화를 진행한다. 우선, 각 행이 시간 포인트이고 각 열이 주식을 나타내는 로그 수익률 데이터 행렬 A가 있다고 하자. 행렬 A에 대해 특잇값 분해를 진행한다.

$$A = USV^T$$

이때 V 는 특이 벡터 행렬로 각 열은 해당 주성분에 대한 각 주식의 로딩(베타)를 나타낸다. 팩터 모델에서 각 주식의 수익률은 요인(팩터)과 이 요인에 대한 민감도(베타)로 설명된다. 여기서 US 는 시간에 따른 주성분(요인)을 나타내고, V 는 각 주식이 이 요인에 어떻게 반응하는지를 나타내는 로딩(베타)를 의미한다. 따라서, V 를 바탕으로 군집화를 진행한다면, 비슷한 위험요소에 노출된, 즉 여러 팩터에 대해 비슷하게 반응하는 주식들끼리 묶을 수 있다. 알고리즘1에서와 같이 고윳값이 1 이상인 주성분만을 채택하지만, 그 수가 15를 넘지 않게 한다. 그 후 DBSCAN, K-Means, Hierarchical 군집화 방법론을 사용한다.

3.2.3 실증분석

앞서 소개한 1번 알고리즘과 2번 알고리즘을 상관계수를 제공한 값 즉 결정계수를 사용하여 평가한다. 상관계수 자체는 두 변수 간의 선형 관계의 강도와 방향을 나타낸다. 시계열 데이터의 군집화 관점에서 결정계수를 사용하여 군집 내 종목들이 얼마나 유사한지를 평가한다. Train 데이터 셋에 알고리즘을 적용하여 군집화를 마친 후 Validation 데이터셋에서 알고리즘의 성능을 평가한다. 알고리즘의 성능을 5회 측정하여 평균한 값은 <표5>와 같다.

R-Squared		Window1	Window2	Window3	Window4	Window5	Average
Algorithm1	Kmeans	0.092587	0.088219	0.081867	0.140964	0.115659	0.103059
	DBSCAN	0.259715	0.02942	0.082085	0.038686	0.220303	0.126042
	Hierarchical	0.07081	0.049294	0.12278	0.080722	0.171922	0.099906
Algorithm2	Kmeans	0.099167	0.083882	0.097379	0.119497	0.10202	0.100789
	DBSCAN	NaN	NaN	NaN	NaN	NaN	NaN
	Hierarchical	0.09686	0.099186	0.10776	0.145029	0.14957	0.119681

<표 2> 군집화 기반 양상블 페어선택 알고리즘 성능비교

1번 알고리즘에서 DBSCAN을 사용했을 때 성능이 가장 좋게 나온 것을 확인할 수 있다. 하지만 군집의 수와 군집내 종목의 수가 너무 적어 본 연구의 목적에 부합하지 않는다. 따라서 두번째로 성능이 좋으며 군집의 개수와 군집내 종목의 개수가 어느정도 보장되는, 2번 알고리즘의 Hierarchical를 사용하여 이후 분석을 진행한다. 5.1 알고리즘1 군집화 결과와, 5.2 알고리즘2 군집화 결과에서 알고리즘별 군집화 결과를 확인할 수 있다.

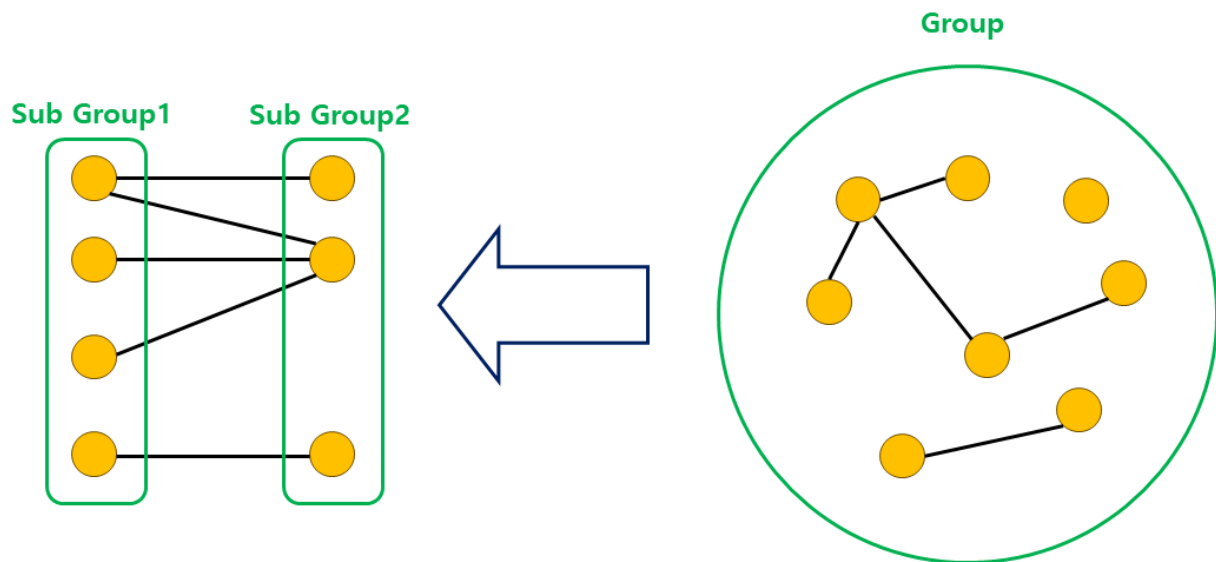
3.3 칼만필터를 활용한 스프레드 동적 추정 알고리즘

3.3.1 one-to-one 페어 선정

군집 내 자산들 간 Engle-Granger 두 단계 공적분 검정(Engle-Granger Two-Step Cointegration Test)을 수행하여 자산 간의 공적분 관계를 식별하고 이를 통해 군집 내 one-to-one 페어를 선정한다. Engle & Granger [6] 검정 순서는 다음과 같다. 첫째, OLS 회귀분석을 통해 두 자산의 시계열 데이터를 사용하여 단순 회귀분석(OLS: Ordinary Least Squares)을 수행한다. 둘째, 회귀 모델의 잔차(residuals)를 계산한다. 셋째, 잔차에 대해 ADF(Augmented Dickey-Fuller) 검정을 사용하여 잔차가 정상성을 갖는지 확인한다.

3.3.2 서브군집 분할 알고리즘

다음은 군집 내 자산들을 두 그룹으로 분할하기 위한 알고리즘이다. <그림4>에서 설명하는 것처럼 공적분 관계를 가지는 one-to-one 페어들을 대상으로 서브군집 분할 알고리즘을 사용하여 자산 쌍을 두 그룹으로 나눈다. 우선, 군집 내 각 주식을 노드로, 공적분 관계를 갖는 노드들을 간선으로 잇는 그래프를 생성한다. 이는 비연결 무방향 그래프이다. 연결된 간선이 없는 노드들은 제거하고 간선이 연결된 노드들을 대상으로 서브군집 분할 알고리즘을 수행한다. 알고리즘의 의사코드(pseudocode)는 <코드1>과 같다.



<그림 4> 서브군집 분할

```

FUNCTION divide_into_two_groups(pairs):
    // Initialize the first pair
    first_pair = pairs[0]['columns']
    group1 = SET()
    group2 = SET()
    // Add the first pair to the two groups
    group1.ADD(first_pair[0])
    group2.ADD(first_pair[1])
    // Iterate over the remaining pairs
    FOR each pair IN pairs[1:]:
        item1, item2 = pair['columns']

        // Check if the items are already in different groups
        IF item1 IN group1 AND item2 IN group2:
            CONTINUE
        ELSE IF item1 IN group2 AND item2 IN group1:
            CONTINUE

        // Assign items to different groups
        ELSE IF item1 IN group1 AND item2 NOT IN group1:
            group2.ADD(item2)
        ELSE IF item1 IN group2 AND item2 NOT IN group2:
            group1.ADD(item2)
        ELSE IF item2 IN group1 AND item1 NOT IN group1:
            group2.ADD(item1)
        ELSE IF item2 IN group2 AND item1 NOT IN group2:
            group1.ADD(item1)

        // If both items are not in any group, assign them to different groups
        ELSE IF (item1 NOT IN group1 AND item2 NOT IN group2) OR (item2 NOT IN group1 AND
item1 NOT IN group2):
            group1.ADD(item1)
            group2.ADD(item2)
        ELSE:
            pairs.REMOVE(pair)
            CONTINUE
    PRINT("Group 1:", LIST(group1))
    PRINT("Group 2:", LIST(group2))
    RETURN group1, group2

```

<코드 1> 서브군집 분할 알고리즘 의사코드

위 코드는 크게 세차례의 과정을 거친다. 1) 초기화 2) 그룹 분할 3) 결과 출력 및 반환

1) 초기화

가) 첫번째 자산 쌍을 이용하여 초기 설정을 한다.

나) 첫 번째 자산을 'group1'에 추가하고, 두 번째 자산을 'group2'에 추가한다.

2) 그룹 분할:

가) 자산 쌍의 첫 번째 자산이 이미 'group1'에 있고, 두 번째 자산이 아직 어떤 그룹에도 속하지 않는다면 이를 'group2'에 추가한다.

나) 반대로, 첫 번째 자산이 'group2'에 있고, 두 번째 자산이 아직 어떤 그룹에도 속하지 않으면 이를 'group1'에 추가한다.

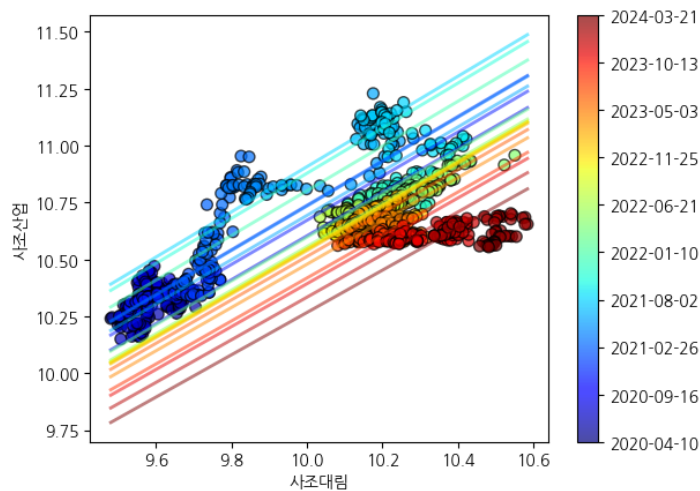
다) 만약 자산 쌍의 두 자산이 어느 그룹에도 속하지 않는다면, 각각 다른 그룹에 추가한다.

3) 결과 출력 및 반환

다만 서브 군집 분할을 적용한 결과가 one-to-one페어라면, 해당 군집은 거래에 사용하지 않는다.

3.3.3 칼만필터를 활용한 스프레드 동적 추정 알고리즘

본 알고리즘은 칼만 필터를 활용하여 공적분 관계를 가지는 자산 쌍의 스프레드를 동적으로 추정하는 과정이다. 자산 쌍의 가격 데이터를 로그 변환하여 독립 변수와 종속 변수를 설정한다. 독립 변수로 관측 행렬을 생성하고, 초기 상태 평균과 공분산, 상태 전이 행렬, 관측 행렬 등을 설정하여 칼만 필터를 초기화한다. 여기서 델타(δ)는 $1e-5$ 로 설정되며, 상태 공분산은 $(\frac{\delta}{1-\delta} \times I)$ 로 계산된다. 필터를 적용하여 동적으로 변하는 회귀 계수를 추정하고, 이 계수를 이용해 각 시점에서의 스프레드를 계산한다. 칼만필터를 통해 스프레드를 갱신하면 <그림5>와 같이 기본 OLS 스프레드에 비해 회귀 계수의 값이 노이즈로 인해 달라지더라도 보정단계를 거쳐 회귀 계수의 값을 보정해줄 수 있다는 장점이 있다.



<그림5> 칼만필터를 통한 회귀계수 보정 시각화 예시

3.3.4 Many-to-Many 페어 스프레드 추정

one-to-one 페어의 스프레드는 칼만필터를 활용한 동적 추정 알고리즘을 활용해 결정한다. one-to-one 페어들을 사용하여 4.3.2의 서브 군집 분할 알고리즘을 활용한 서브 군집에 따라 many-to-many 페어를 결정한다. Many-to-many 페어를 구성하는 각 one-to-one 페어의 비중은 다음과 같이 결정된다. 페어트레이딩에서는 스프레드의 평균회귀를 기대하며 베팅하기 때문에 페어의 공적분 관계가 유지되는 것이 중요하다. 베팅 사이즈를 결정하기 위해, 각 페어 Engle-

Granger 두 단계 공적분 검정의 ADF-test에서 결정된 p-value를 활용한다.

- 1) 먼저, 각 p-value의 역수에 로그를 씌운 값을 계산한다: $\log\left(\frac{1}{p_i}\right)$
- 2) 모든 항목에 대해 위 값을 합산한다: $\text{Total} = \sum_{i=1}^n \log\left(\frac{1}{p_i}\right)$
- 3) 각 항목의 비중은 다음과 같이 계산된다: $\text{weight}_i = \frac{\log\left(\frac{1}{p_i}\right)}{\text{Total}}$
- 4) $\sum_{i=1}^n \text{weight}_i = 1$

각 one-to-one페어에 앞서구한 weight를 적용시켜 만든 many-to-many 페어에 대해 ADF-test를 진행하고, 정상성이 있다고 판단되면 해당 many-to-many 페어에 대해 거래를 진행한다.

3.3.5 거래 전략

- 1) 최대 보유기간 설정: 반감기
- 1) 손절 기준: 3표준편차
- 2) 매수 시그널: 2 표준편차 이탈 후 진입시
- 3) 매도 시그널: 평균 \pm (2 표준편차 \times 0.1)

거래전략의 기본은 Herlemont, D[5]에서 밝힌 표준편차를 통한 스프레드 매매에서 따왔다.

1)에서 반감기는 OU process 라는 평균회귀 확률 과정 방정식으로부터 계산할 수 있다.

$$dy_t = \theta(\mu - y_t)dt + \sigma dW_t$$

다음과 같은 OU process 공식을 적절히 전개하면 다음과 같은 식이 나온다.

$$X_t = X_0 + \mu(e^{-\lambda t} - 1) + \sigma \int_0^t e^{-\lambda s} dW_s$$

여기서 W_s 는 기하브라운 운동에서의 Winner process를 의미하므로 평균은 0에 수렴한다. 이를 반영하여 정리를 하면 결론적으로 공식은 다음과 같이 유도된다.

$$y_t = y_0 e^{-\lambda t} + \mu(1 - e^{-\lambda t})$$

따라서 반감기 T (즉, y_t 가 절반이 되는 시점)는 다음과 같이 계산된다:

$$T = \frac{\ln(2)}{\ln(1/\phi)}$$

우리가 최대 보유기간을 반감기로 삼는 이유는 평균회귀의 초기 단계에서 가장 빠른 속도로 변화가 일어나므로, 이 기간 동안의 투자는 더 높은 효율성(즉, 시간 대비 수익)을 제공하기 때문

이다.

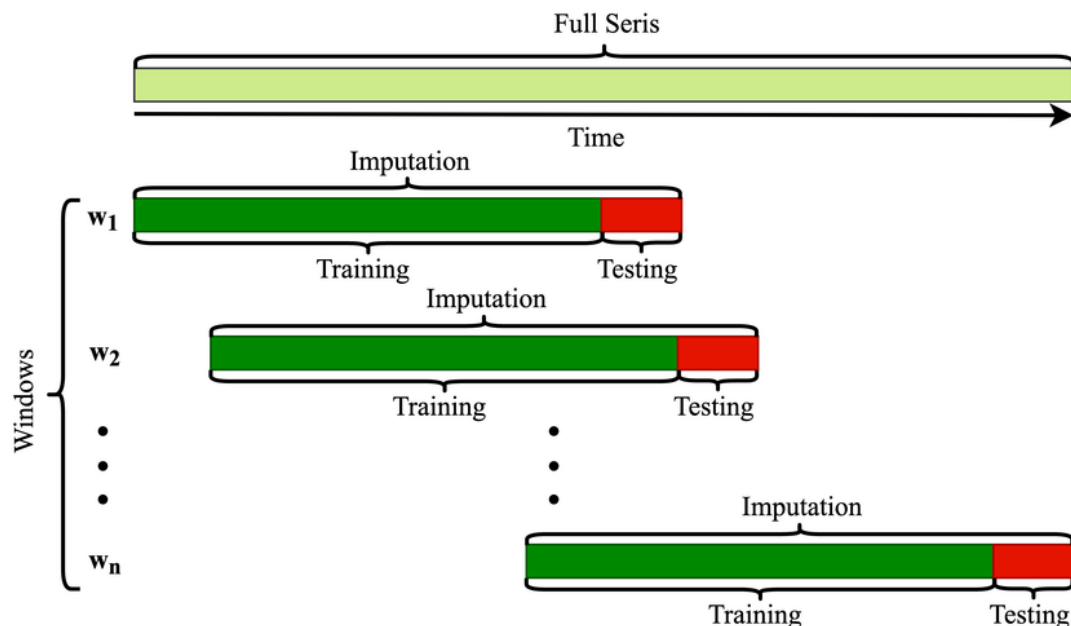
3.3.6 백테스트

1) 벤치마크 설정

페어 트레이딩 전략은 시장 중립 전략이고, 포지션이 잡히지 않을 때는 현금자산으로 보유한다. 따라서, 전략의 성과 평가 시 벤치마크로서 무위험 수익률을 설정하는 대신 절대 기준 수익률을 0으로 설정한다. 이는 시장의 상승 또는 하락과 관계없이 전략의 순수한 알파(α)를 측정하고, 시장의 방향성에 대한 노출을 최소화한 상태에서의 성과를 평가하는 데 중점을 둔다.

2) 롤링 윈도우 분석(Rolling Window Analysis)

롤링 윈도우 분석(Rolling Window Analysis)은 슬라이딩 윈도우(sliding window) 방식을 사용하여 전략의 성과를 평가하는 방법이다. <그림6>과 같이 일정 크기의 시간 구간(윈도우)을 설정하고, 이 윈도우를 일정 간격으로 이동시키면서 각 구간마다 전략의 성과를 평가한다. 본 연구에서 사용된 첫 번째 윈도우의 훈련 데이터는 1999년 5월 14일부터 2016년 1월 4일까지이고, 테스트 데이터는 2016년 1월 5일부터 2020년 3월 17일까지이다. 두 번째 윈도우부터는 256 영업일만 큼 일정 간격으로 이동하면서 각 구간마다 전략의 성과를 평가한다.



<그림6> Rolling Window Analysis[11]

3) 거래비용

롱(Long) 포지션 거래비용: 0.125%, 숏(Short) 포지션 거래비용: 1.53%를 고려하여 백테스트를 진행하였다.

4) 벡테스트 결과

모든 Window에서 벤치마크를 상회하는 모습을 볼 수 있다.

Window	1	2	3	4	5
구간 수익률(1030영업일: 약 4년, 소수 3째 자리에서 반올림)	0.342%	1.36%	5.06%	1.37%	1.77%
CAGR(Compound Annual Growth Rate)	0.0854%	0.338%	1.2315%	0.341%	0.440%

<표3> Rolling Window Analysis

IV. 결론

본 연구는 페어 트레이딩 전략의 효율성을 극대화하기 위해 페어 선택 및 거래 전략 방식을 개선하였다. 페어 선택 부분에서는 Simão Moraes Sarmento[9]에서 제안된 군집화 기반 양상블 페어 선택 알고리즘을 팩터 모델의 관점에서 개선하였다. 개선된 알고리즘과 기존 알고리즘을 비교한 결과, 계층적 군집화를 적용한 개선된 알고리즘이 가장 효과적인 것으로 나타났다. 또한, 거래 전략 방식에서 Wang[7]이 제안한 MMPT의 one-to-one 페어 일부가 소실되는 문제를 서브군집 분할 알고리즘을 사용하여 해결하였다. 칼만 필터를 통해 스프레드를 동적으로 추정함으로써 효율적인 MMPT를 구현할 수 있었다. 테스트 결과, 각 윈도우에서 CAGR이 0.0854%, 0.338%, 1.2315%, 0.341%, 0.440%로 전반적으로 벤치마크를 상회하였다. 특히 2016년부터 2024년까지는 코로나19 팬데믹 상황, 금리의 급격한 변화, 러-우 전쟁 등 다양한 대내외적 요인으로 시장 상황이 급변하는 시기였다. 본 연구에서 사용한 전략이 이러한 변동성에도 불구하고 안정적으로 수익을 낸다는 점은 주목할 만하다. 그러나 본 연구에는 몇 가지 한계가 있다. 첫째, 연구 기간 내 신규 상장 및 상장 폐지된 종목을 고려하지 않았으므로 미래 참조 편향 및 생존 편향이 발생할 우려가 있다. 둘째, 페어 선택에서 PCA(주성분 분석)은 선형 관계만을 분석하므로 비선형 관계를 포착하지 못한다는 단점이 있다. 셋째, 더 진보된 군집화 알고리즘을 사용하여 페어 선택 알고리즘의 성능을 향상시킬 여지가 있다. 넷째, 칼만 필터는 동적 선형 모델이므로 확장 칼만 필터를 통해 비선형 관계를 포착할 수 있는 가능성이 있다. 이러한 한계를 보완하기 위해 향후 연구에서는 신규 상장 및 상장 폐지된 종목을 포함하는 데이터 세트를 사용할 필요가 있으며, 비선형 관계를 분석할 수 있는 기법을 도입하는 것이 필요하다. 더불어, 새로운 군집화 알고리즘과 확장 칼만 필터를 활용한 연구를 통해 페어 트레이딩 전략의 효율성을 더욱 향상시킬 수 있을 것이다.

V. 부록

5.1. 알고리즘1 군집화 결과

Algorithm	Cluster Method	Cluster	Stocks
algorithm1 - window1	KMeans	1	LG, HD 한국조선해양, 삼성중공업, 현대제철, HD 현대미포, 동국홀딩스, 한진중공업홀딩스
	KMeans	2	삼성전자, SK 하이닉스, 삼성 SDI, 삼성전기
	KMeans	3	동원산업, 사조대림, 사조산업, 신라교역, 사조오양, 한성기업, 동원수산
	KMeans	4	삼성화재, 기업은행, 코리안리
	KMeans	5	현대건설, 삼부토건, 코오롱글로벌, 계룡건설, 동부건설, HL D&I
	KMeans	6	삼화콘덴서, 삼화전기, 삼화전자, 성문전자
	KMeans	7	현대차, 기아, 현대모비스, 한국엔컴퍼니
	KMeans	8	동서, 콜마홀딩스, 태경케미칼, 무학, KH 필룩스, 한국내화, 황금에스티, 인팩, 부국철강
	KMeans	9	CJ, 에스원, 제일기획, 신세계, 오리온홀딩스, 빙그레, SBS
	KMeans	10	태광산업, 고려제강, BYC, 대한화섬, 비비안
	KMeans	11	SK 가스, E1, 삼천리, 한국셀석유, 예스코홀딩스, 경동인베스트
	KMeans	12	유한양행, 한미사이언스, 대웅, 녹십자홀딩스, 동아쏘시오홀딩스, 환인제약
	KMeans	13	POSCO 홀딩스, 한국전력
	KMeans	14	고려아연, S-Oil, LS, 롯데케미칼, 금호석유, KCC, OCI 홀딩스
	KMeans	15	SK 텔레콤, KT
	KMeans	16	쌍용 C&E, 한일현대시멘트, 동양, 성신양회, 진흥기업, 까뮤이앤씨, 남광토건, 범양건영
	KMeans	17	GS 건설, DL, HDC

	KMeans	18	한화솔루션, SKC, 한화, 효성, 롯데정밀화학, SK 디스커버리, 풍산홀딩스, 코오롱
	KMeans	19	에스엘, 화신, 대원강업, 서연, 에스제이지세종, 디와이
	KMeans	20	신영증권, 부국증권, 유화증권, 상상인증권
	KMeans	21	보령, JW 중외제약, 부광약품, 광동제약, 종근당홀딩스, 삼진제약, 일양약품, 일성아이에스, 동화약품
	KMeans	22	한온시스템, 한국단자, 삼영전자, 대덕, 한국전자홀딩스
	KMeans	23	DB 손해보험, 현대해상, 롯데손해보험, 한화손해보험, 제주은행, 흥국화재
	KMeans	24	만호제강, 성창기업지주, 방림, SG 글로벌, 대한방직, SUN&L
	DBSCAN	1	SK 텔레콤, KT
	Hierarchical	1	삼성전자, SK 하이닉스, POSCO 홀딩스, 삼성 SDI, 한국전력, 삼성전기, SK 텔레콤, KT
	Hierarchical	2	S-Oil, 영풍, 태광산업, 고려제강, 송원산업, 경방, KCTC, 크라운해태홀딩스, 신영와코루, 카프로
	Hierarchical	3	BYC, 만호제강, 대한화섬, 성창기업지주, 방림, 대한방직, SUN&L, 비비안
	Hierarchical	4	현대차, 기아, 현대모비스, 한온시스템, 한국엔컴퍼니
	Hierarchical	5	삼화콘덴서, 삼화전기, 삼화전자, 성문전자
algorithm1 - window2	KMeans	1	KG 모빌리티, 세아베스틸지주, 제주은행, LS 네트워크, GS 글로벌, 신원, 한신공영, 일성건설
	KMeans	2	CJ, 제일기획, 오뚜기, 오리온홀딩스, 대상, 빙그레, 삼양홀딩스, 한섬, SBS
	KMeans	3	미래에셋증권, NH 투자증권, 삼성증권, 대신증권
	KMeans	4	현대차, 기아, 현대모비스
	KMeans	5	에스엘, 화신, 서연, 에스제이지세종, 디와이
	KMeans	6	아모레 G, 롯데지주, 농심, 롯데칠성, 영풍, 퍼시스, 한일홀딩스, 신도리코, 남양유업

	KMeans	7	동원산업, 사조대림, 사조산업, 신라교역, 사조오양, 한성기업, 동원수산
	KMeans	8	세아제강지주, 극동유화, 한국특강, 한일철강, 부국철강
	KMeans	9	SK 하이닉스, 카카오, 동원시스템즈, 디아이, 신성이엔지, 와이투솔루션, 콤텍시스템, 미래산업
	KMeans	10	DN 오토모티브, 태광산업, 고려제강, 아세아, KISCO 홀딩스, CR 홀딩스, 대한제분, KPX 케미칼, 하이트진로홀딩스, 경방
	KMeans	11	대한항공, LS ELECTRIC, 호텔신라, 롯데정밀화학, 아시아나항공, SK 디스커버리, 한진, 코오롱, 한솔홀딩스
	KMeans	12	삼화콘덴서, 삼화전기, 삼화전자, 성문전자
	KMeans	13	쌍용 C&E, 한일현대시멘트, 성신양회
	KMeans	14	삼성 E&A, 현대건설, GS 건설, DL, HDC
	KMeans	15	SK 가스, E1, 삼천리, 한국셀석유, 예스코홀딩스, 신대양제지, 경농, 태경산업, 경동인베스트, 미창석유
	KMeans	16	삼성화재, 기업은행, 코리안리
	KMeans	17	DB 손해보험, 현대해상, 롯데손해보험, 한화손해보험, 흥국화재
	KMeans	18	DI 동일, BYC, 만호제강, 대한화섬, 성창기업지주, 방림, 대한방직
	KMeans	19	삼성전자, 삼성 SDI, 삼성전기, 한화에어로스페이스, 대덕
	KMeans	20	한국전력, KT&G, SK 텔레콤, KT
	KMeans	21	SNT 다이내믹스, 에이프로젠, 현대비앤지스틸, SIMPAC, STX, 동원금속, 대양금속, 케이비아이동국실업
	KMeans	22	휴스틸, NI 스틸, 동양철관, 문배철강
	KMeans	23	삼부토건, 동양, 코오롱글로벌, 진흥기업, 계룡건설, 동부건설, HL D&I, 까뮤이앤씨, 남광토건, 범양건영
	KMeans	24	유한양행, 한미사이언스, 녹십자, 대웅, 녹십자홀딩스, 동아쏘시오홀딩스, 종근당홀딩스
	KMeans	25	한온시스템, 한국가스공사, 에스원, 신세계
	KMeans	26	POSCO 홀딩스, HD 한국조선해양, S-Oil, 현대제철, 동국홀딩스

	KMeans	27	영진약품, 진원생명과학, 삼성제약, 에이프로젠바이오로지스, CJ 씨푸드, 팜젠사이언스, 오리엔트바이오
	KMeans	28	한국엔컴퍼니, 넥센타이어, 한국단자, 대원강업, 모토닉, 한국무브넥스, 인지컨트롤스, 화승코퍼레이션, 인팩, 유성기업
	DBSCAN	1	S-Oil, 동원산업, 사조대림, 사조산업, 신라교역, 사조오양, 한성기업, 동원수산
	Hierarchical	1	삼성전자, 현대차, 기아, 삼성 SDI, 현대모비스, 삼성전기
	Hierarchical	2	에스엘, 화신, 대원강업, 모토닉, 서연, 에스제이지세종, 한국무브넥스, 인지컨트롤스, 화승코퍼레이션, 유성기업
	Hierarchical	3	삼화콘덴서, 삼화전기, 삼화전자, 성문전자
	Hierarchical	4	삼성화재, 기업은행, DB 손해보험, 현대해상, 코리안리, 롯데손해보험, 한화손해보험, 흥국화재
	Hierarchical	5	한국전력, KT&G, SK 텔레콤, KT, 한국가스공사
	Hierarchical	6	HLB 글로벌, 휴스틸, NI 스틸, 광명전기, 동양철관, 대양금속, 문배철강
	Hierarchical	7	동원산업, 사조대림, 사조산업, 신라교역, 사조오양, 한성기업, 동원수산
	Hierarchical	8	DI 동일, BYC, 만호제강, 대한화섬, 성창기업지주, 방림, 대한방직
algorithm1 - window3	KMeans	1	POSCO 홀딩스, 두산에너지빌리티, 한화오션, HD 한국조선해양, 삼성중공업, 현대제철, HD 현대미포, HD 현대인프라코어, 동국홀딩스
	KMeans	2	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	KMeans	3	일성아이에스, 삼일제약, 한독, 제일파마홀딩스, 일동홀딩스
	KMeans	4	코오롱글로벌, 계룡건설, 동부건설, 화성산업, HL D&I
	KMeans	5	NI 스틸, 동양철관, 문배철강
	KMeans	6	KCC, 현대엘리베이, SK 네트웍스, 대한유화, 세아제강지주, 세아베스틸지주, 풍산홀딩스
	KMeans	7	한국전력, KT&G, SK 텔레콤, KT, 한국가스공사
	KMeans	8	삼화콘덴서, 삼화전기, 삼화전자

	KMeans	9	DB 하이텍, 송원산업, 현대비엔지스틸, 이수화학, AK 홀딩스, 대창, 유니온, 서원, 이건산업, 카프로
	KMeans	10	현대차, 기아, 현대모비스
	KMeans	11	카카오, 엔씨소프트, 코웨이, LG 유플러스, 영원무역홀딩스, 현대지에프홀딩스, 제주은행, 흥국화재, HS 애드, 대구백화점
	KMeans	12	삼부토건, 동양, 진흥기업, 한신공영, 일성건설, 까뮤이앤씨, 남광토건, 범양건영
	KMeans	13	한온시스템, 아모레 G, 롯데지주, 농심, 에스원, 신세계, 롯데칠성, 신도리코, 남양유업
	KMeans	14	삼성전자, 삼성 SDI, 삼성전기, 한화에어로스페이스, 한국엔컴퍼니, 대덕
	KMeans	15	SK 가스, E1, 삼천리, 한국셀석유, 예스코홀딩스, 신대양제지, 경농, 미창석유, 극동유화, 한국수출포장
	KMeans	16	LG 화학, LG, S-Oil, 한화솔루션, 롯데케미칼
	KMeans	17	삼성화재, 기업은행, DB 손해보험, 현대해상, 코리안리, 한화손해보험
	KMeans	18	현대건설, 대우건설, GS 건설, DL, HDC
	KMeans	19	영진약품, 진원생명과학, 삼성제약, 에이프로젠바이오로직스, 팜젠사이언스, 유유제약, 오리엔트바이오
	KMeans	20	신영증권, 부국증권, 유화증권, 상상인증권
	KMeans	21	포스코인터내셔널, 대한전선, 대한해운, KG 스틸, STX, 금호건설, 유수홀딩스, 태영건설
	KMeans	22	쌍용 C&E, 한일홀딩스, 한일현대시멘트, 성신양회
	KMeans	23	유한양행, 한미사이언스, 녹십자, 녹십자홀딩스, 동아쏘시오홀딩스
	DBSCAN	1	S-Oil, 삼화콘덴서, 삼화전기, 삼화전자
	Hierarchical	1	삼화콘덴서, 삼화전기, 삼화전자, 성문전자
	Hierarchical	2	현대차, 기아, 현대모비스
	Hierarchical	3	한국전력, KT&G, SK 텔레콤, KT, 한국가스공사
	Hierarchical	4	현대건설, 대우건설, GS 건설, DL, HDC
	Hierarchical	5	BYC, 만호제강, 대한화섬, 성창기업지주, 방림, 대한방직
	Hierarchical	6	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산

	Hierarchical	7	NI 스틸, 광명전기, 동양철관, 문배철강, 선도전기
algorithm1 - window4	KMeans	1	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	KMeans	2	미래에셋증권, NH 투자증권, 삼성증권, 대신증권
	KMeans	3	LG 생활건강, CJ, 아모레 G, 농심, 오뚜기, 롯데칠성, 오리온홀딩스, 대상, 동원 F&B, 삼양홀딩스
	KMeans	4	한일현대시멘트, 성신양회, 경농, HL D&I, 일성건설
	KMeans	5	삼성전자, SK 하이닉스, 삼성 SDI, LG 전자, 삼성전기
	KMeans	6	조광피혁, 삼화페인트, 신세계 I&C, 강남제비스코, 삼양통상, 성창기업지주, 디씨엠, 방림, 까뮤이앤씨, 대한방직
	KMeans	7	대원전선, 신원, 광명전기, 한국내화, 동양철관, 일신석재, 헤인, 조비, 선도전기, 인디에프
	KMeans	8	코웨이, LG 유플러스, 강원랜드, 에스원, 제일기획, 신세계, 빙그레, 한섬, SBS
	KMeans	9	세아제강지주, 가온전선, 휴스틸, 태경산업, 황금에스티, 화성산업, 유니온, 신세계건설, 이건산업
	KMeans	10	현대비앤지스틸, 대창, 서원
	KMeans	11	삼화콘덴서, 삼화전기, 삼화전자
	KMeans	12	삼성화재, DB 손해보험, 현대해상, 코리안리
	KMeans	13	POSCO 홀딩스, 현대제철, 동국홀딩스
	KMeans	14	한국전력, KT&G, SK 텔레콤, KT, 롯데지주, 한국가스공사, 신도리코
	KMeans	15	현대차, 기아, 현대모비스
	KMeans	16	한온시스템, 한국엔컴퍼니, 넥센타이어, 한국단자
	KMeans	17	유한양행, 한미사이언스, 녹십자, 대웅, 녹십자홀딩스, 동아쏘시오홀딩스, 종근당홀딩스, 환인제약, 한독
	KMeans	18	NI 스틸, 부국철강, 문배철강
	KMeans	19	LG 화학, LG, 고려아연, S-Oil, 한화솔루션, 롯데케미칼, OCI 홀딩스
	KMeans	20	HMM, 쌍용 C&E, 현대엘리베이, 남해화학, 현대코퍼레이션
	KMeans	21	영진약품, 일성아이에스, 삼일제약, 진원생명과학, 제일파마홀딩스, 삼성제약, 에이프로젠바이오텍스, 팜젠사이언스, 오리엔트바이오
	KMeans	22	신한지주, 기업은행

	DBSCAN	1	S-Oil, 동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	Hierarchical	1	코오롱글로벌, 계룡건설, 동부건설, 화성산업, 태영건설, 신세계건설, 한신공영, HL D&I
	Hierarchical	2	신한지주, 삼성화재, 기업은행, DB 손해보험, LG 유플러스, 현대해상, 코리안리
	Hierarchical	3	삼화콘덴서, 삼화전기, 삼화전자
	Hierarchical	4	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
algorithm1 - window5	KMeans	1	LG 생활건강, CJ, 아모레 G, 오뚜기, 오리온홀딩스, 대상, 동원 F&B, 삼양홀딩스
	KMeans	2	한올바이오파마, JW 중외제약, 부광약품, 영진약품, 광동제약, 삼진제약, 일양약품, 동화약품
	KMeans	3	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	KMeans	4	포스코퓨처엠, 엘앤에프, 코스모신소재, 코스모화학, 송원산업, 삼익 THK, 일진디스플레이
	KMeans	5	세방전지, KG 모빌리티, 넥센타이어, SNT 모티브, SNT 다이내믹스, SNT 홀딩스, 디와이
	KMeans	6	신영증권, 부국증권, 다올투자증권, 유화증권, 상상인증권
	KMeans	7	CJ 대한통운, 동서, 한샘, NICE 평가정보, 풀무원, 한세에스 24 홀딩스, 벽산, 샘표, 크라운해태홀딩스, 티웨이홀딩스
	KMeans	8	S-Oil, 대한유화, TKG 휴켄스, 세아베스틸지주, 국도화학, 이수화학, 카프로
	KMeans	9	한온시스템, 에스엘, 화신, 서연, 에스제이지세종
	KMeans	10	두산에너지빌리티, 한화오션, HD 한국조선해양, 삼성중공업, 삼성 E&A, HD 현대미포, HD 현대인프라코어
	KMeans	11	삼성전자, SK 하이닉스, 삼성 SDI, LG 전자, LG, 삼성전기
	KMeans	12	조광피혁, BYC, 삼양통상, 대한화섬, 디씨엠, 방림, 신영와코루, 까뮤이앤씨, 대한방직
	KMeans	13	신한지주, 삼성화재, 기업은행, DB 손해보험, 현대해상
	KMeans	14	현대건설, 대우건설, GS 건설, DL, HDC

	KMeans	15	NI 스틸, 동양철관, DSR 제강, 문배철강
	KMeans	16	유한양행, 한미사이언스, 녹십자, 대응제약, 녹십자홀딩스, 동아쏘시오홀딩스, 종근당홀딩스
	KMeans	17	현대차, 기아, 현대모비스
	KMeans	18	삼부토건, 코오롱글로벌, 금호건설, 계룡건설, 동부건설, 화성산업, 태영건설, 신세계건설, 한신공영, HL D&I
	KMeans	19	POSCO 홀딩스, 고려아연, 현대제철, 세아제강지주, 현대비앤지스틸, 동국홀딩스
	KMeans	20	파미셀, 일성아이에스, 진원생명과학, 에이프로젠바이오로직스, 팜젠사이언스, 보락, 오리엔트바이오
	KMeans	21	삼화콘덴서, 삼화전기, 삼화전자
	KMeans	22	LG 화학, LS ELECTRIC, 한화솔루션, 롯데케미칼, SKC, 금호석유, 한화, OCI 홀딩스, 롯데정밀화학
	KMeans	23	포스코스틸리온, 이구산업, 대창, 서원, 부국철강
	KMeans	24	HMM, 쌍용 C&E, 현대엘리베이, 남해화학, 현대코퍼레이션, 성신양회
	DBSCAN	1	현대차, 기아, 현대모비스
	DBSCAN	2	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	Hierarchical	1	신한지주, 기업은행, DB 손해보험, 현대해상
	Hierarchical	2	NI 스틸, 동양철관, 하이스틸, 부국철강, DSR 제강, 문배철강
	Hierarchical	3	삼성전자, SK 하이닉스, 삼성 SDI, LG 전자, 삼성전기
	Hierarchical	4	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	Hierarchical	5	현대차, 기아, 현대모비스

5.2. 알고리즘2 군집화 결과

Algorithm	Cluster Method	Cluster	Stocks
-----------	----------------	---------	--------

algorithm2 - window1	KMeans	1	대한항공, LS ELECTRIC, 삼성 E&A, 한화솔루션, SKC, 한화, 롯데정밀화학, LX 인터내셔널, SK 디스커버리, 코오롱
	KMeans	2	동원산업, 사조대림, 사조산업, 신라교역, 사조오양, 한성기업, 동원수산
	KMeans	3	POSCO 홀딩스, LG, 한화에어로스페이스, 신세계
	KMeans	4	태광산업, 고려제강, BYC, 대한화섬, 비비안
	KMeans	5	삼성전자, SK 하이닉스, 삼성 SDI, 한국전력, 삼성전기, SK 텔레콤, KT
	KMeans	6	현대차, 기아, 현대모비스, 한온시스템, 한국엔컴퍼니
	KMeans	7	삼화콘덴서, 삼화전기, 삼화전자, 성문전자
	KMeans	8	삼성화재, 기업은행, DB 손해보험, 현대해상, 코리안리, 롯데손해보험, 한화손해보험, 제주은행, 흥국화재
	KMeans	9	콜마홀딩스, KH 필룩스, 한국내화
	KMeans	10	유한양행, 한미사이언스, 대웅, 녹십자홀딩스, JW 증외제약, 동아쏘시오홀딩스, 부광약품, 종근당홀딩스, 일양약품
	KMeans	11	만호제강, 성창기업지주, 방림, 대한방직, SUN&L
	KMeans	12	에스엘, 화신, 서연, 에스제이지세종
	KMeans	13	코웨이, 동서, 태경케미컬, 무학, 황금에스티, 인팩, 부국철강
algorithm2 - window2	Hierarchical	1	삼성전자, SK 하이닉스, POSCO 홀딩스, 삼성 SDI, 한국전력, 삼성전기, SK 텔레콤, KT
	Hierarchical	2	코웨이, 동서, 콜마홀딩스, 태경케미컬, 에스제이지세종, KH 필룩스, 한국내화, 황금에스티, 인팩, 부국철강
	Hierarchical	3	태광산업, BYC, 만호제강, 대한화섬, 성창기업지주, 방림, 대한방직, SUN&L, 비비안
	Hierarchical	4	동원산업, 사조대림, 사조산업, 신라교역, 사조오양, 한성기업, 동원수산
	Hierarchical	5	삼성화재, 기업은행, DB 손해보험, 현대해상, 코리안리, 롯데손해보험, 한화손해보험, 제주은행, 흥국화재
	Hierarchical	6	현대차, 기아, 현대모비스
	Hierarchical	7	삼화콘덴서, 삼화전기, 삼화전자, 성문전자
algorithm2 - window2	KMeans	1	동원산업, 사조대림, 사조산업, 신라교역, 사조오양, 한성기업, 동원수산

	KMeans	2	SK 하이닉스, 삼성 SDI, LG, 삼성전기, 한화에어로스페이스, 신세계, SK 네트워크, 삼성전자, 대덕, 한국전자홀딩스
	KMeans	3	현대비앤지스틸, 휴스틸, NI 스틸, 동양철관, 문배철강
	KMeans	4	DI 동일, BYC, 만호제강, 대한화섬, 성창기업지주, 방림, 대한방직
	KMeans	5	삼성전자, POSCO 홀딩스, 한국전력, KT&G, SK 텔레콤, KT, 한국가스공사
	KMeans	6	삼화콘덴서, 삼화전기, 삼화전자, 성문전자
	KMeans	7	삼성화재, 기업은행, DB 손해보험, 현대해상, 코리안리, 롯데손해보험, 한화손해보험, 흥국화재
	KMeans	8	현대차, 기아, 현대모비스
	KMeans	9	삼성 E&A, GS 건설, DL, HDC, HL D&I
	Hierarchical	1	삼화콘덴서, 삼화전기, 삼화전자, 성문전자
	Hierarchical	2	삼성화재, 기업은행, DB 손해보험, 현대해상, 코리안리, 롯데손해보험, 한화손해보험, 제주은행, 흥국화재
	Hierarchical	3	동원산업, 사조대림, 사조산업, 신라교역, 사조오양, 한성기업, 동원수산
	Hierarchical	4	한국전력, KT&G, SK 텔레콤, KT, 한국가스공사
	Hierarchical	5	태광산업, BYC, 만호제강, 대한화섬, 성창기업지주, 방림, 대한방직
	Hierarchical	6	현대차, 기아, 현대모비스
algorithm2 - window3	KMeans	1	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	KMeans	2	두산에너지빌리티, 한화오션, HD 한국조선해양, 삼성중공업, 삼성 E&A, HD 현대미포, 두산, HD 현대인프라코어, 한진중공업홀딩스
	KMeans	3	한온시스템, 에스엘, 한국엔컴퍼니, 넥센타이어, 화신, 서연, 에스제이지세종, 디와이
	KMeans	4	한국전력, KT&G, SK 텔레콤, KT, 한국가스공사
	KMeans	5	삼화콘덴서, 삼화전기, 삼화전자
	KMeans	6	삼성화재, 기업은행, DB 손해보험, 현대해상, 코리안리, 롯데손해보험, 한화손해보험, 흥국화재
	KMeans	7	BYC, 만호제강, 대한화섬, 성창기업지주, 방림, 대한방직

	KMeans	8	현대건설, 대우건설, GS 건설, DL, HDC, 계룡건설, HL D&I
	KMeans	9	SK 가스, E1, 한국셀석유, 태경산업, 미창석유
	KMeans	10	현대차, 기아, 현대모비스
	Hierarchical	1	DB 손해보험, 현대해상, 롯데손해보험, 한화손해보험, 흥국화재, 진양산업, 진양폴리, 진양화학
	Hierarchical	2	BYC, 만호제강, 대한화섬, 성창기업지주, 방림, 깨끗한나라, 대한방직
	Hierarchical	3	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	Hierarchical	4	한국전력, KT&G, SK 텔레콤, KT, 한국가스공사
	Hierarchical	5	삼화콘덴서, 삼화전기, 삼화전자
	Hierarchical	6	현대차, 기아, 현대모비스
algorithm2 - window4	KMeans	1	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	KMeans	2	신영증권, 부국증권, 한양증권, 유화증권, 상상인증권
	KMeans	3	삼화콘덴서, 삼화전기, 삼화전자, 성문전자
	KMeans	4	신한지주, 삼성화재, 한국전력, KT&G, SK 텔레콤, 기업은행, KT, 강원랜드, 한국가스공사, 하이트진로홀딩스
	KMeans	5	현대차, 기아, 현대모비스, 에스엘, 서연
	KMeans	6	두산에너지빌리티, 한화오션, HD 한국조선해양, 삼성중공업, S-Oil, HD 현대미포
	KMeans	7	카카오, 대한항공, 엔씨소프트, 영원무역홀딩스, 아시아나항공, 현대지에프홀딩스, NICE 평가정보, 한진, 광주신세계, 대구백화점
	KMeans	8	삼성전자, SK 하이닉스, 삼성 SDI, LG 전자, 삼성전기
	KMeans	9	현대건설, 대우건설, GS 건설, DL, HDC
	KMeans	10	현대비앤지스틸, 대창, NI 스틸, 동양철관, 서원, 한일철강, 부국철강, 문배철강
	KMeans	11	코웨이, LG 유플러스, 제일기획, 코리안리, 빙그레, 한섬, SBS, 한솔홀딩스
	KMeans	12	POSCO 홀딩스, 현대제철, 동국홀딩스
	KMeans	13	유한양행, 한미사이언스, 녹십자, 대웅, 녹십자홀딩스, 동아쏘시오홀딩스, 종근당홀딩스, 환인제약, 한독

	Hierarchical	1	신한지주, 삼성화재, 한국전력, KT&G, SK 텔레콤, 기업은행, KT, LG 유플러스, 한국가스공사, 코리안리
	Hierarchical	2	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	Hierarchical	3	삼화콘덴서, 삼화전기, 삼화전자
	Hierarchical	4	삼성전자, SK 하이닉스, 삼성 SDI, LG 전자, 삼성전기
	Hierarchical	5	현대차, 기아, 현대모비스
algorithm2 - window5	KMeans	1	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	KMeans	2	한온시스템, 에스엘, 화신, 서연, 에스제이지세종
	KMeans	3	휴스틸, NI 스틸, 동양철관, 대호에이엘, 하이스틸, 한일철강, 부국철강, DSR 제강, 문배철강
	KMeans	4	신한지주, 삼성화재, 한국전력, SK 텔레콤, 기업은행, KT, LG 유플러스, 강원랜드, 한국가스공사
	KMeans	5	삼성전자, SK 하이닉스, 삼성 SDI, LG 전자, LG, 삼성전기
	KMeans	6	현대차, 기아, 현대모비스, 한국엔컴퍼니
	KMeans	7	삼화콘덴서, 삼화전기, 삼화전자
	Hierarchical	1	신한지주, 삼성화재, 한국전력, SK 텔레콤, 기업은행, KT, LG 유플러스, 강원랜드, 한국가스공사, 코리안리
	Hierarchical	2	휴스틸, NI 스틸, 동양철관, 대호에이엘, 하이스틸, 한일철강, 부국철강, DSR 제강, 문배철강
	Hierarchical	3	삼화콘덴서, 삼화전기, 삼화전자
	Hierarchical	4	동원산업, 사조대림, 사조산업, 신라교역, CJ 씨푸드, 사조오양, 한성기업, 동원수산
	Hierarchical	5	현대차, 기아, 현대모비스
	Hierarchical	6	삼성전자, SK 하이닉스, 삼성 SDI, LG 전자, 삼성전기

5.3 전략 벡테스트

Cluster		수익률																	
Window1	Cluster1	0.78%	Many to Many Pair		한국전력	KT													
			Pairs		SK텔레콤	SK텔레콤													
			Weights		0.5139345	0.4860655													
			Many to Many Pair		부국셀강	황금에스티	한국내화	태경케미컬											
	Cluster2	-1.15%	Pairs		황금에스티	한국내화	태경케미컬	KH 필룩스	태경케미컬	태경케미컬									
			Weights		0.21140891	0.19251307	0.16369552	0.14870515	0.14186185	0.14181549									
			Many to Many Pair		대한화선	태광산업	BYC	성창기업지주	대한화선	대한화선	성창기업지주								
			Pairs		성창기업지주	BYC	대한방직	민호제강	SUN&L	비비안	대한화선	성창기업지주	태광산업						
	Cluster3	0.15%	Weights		0.19489915	0.15609244	0.13560649	0.11117234	0.10672954	0.10325737	0.0999413	0.09224854							
			Many to Many Pair		시조대릴	신라교역	한성기업	시조산업	동원수산										
Pairs			시조오양	시조대릴	시조산업	신라교역	시조대릴	시조대릴	동원수산										
Weights			0.23729024	0.2051397	0.18970707	0.18429979	0.18356321												
Window2	Cluster4	1.93%	Many to Many Pair		한화손해보험	롯데손해보험	계주는행	한국화재	DB손해보험	기업은행	기업은행	한화손해보험	기업은행						
			Pairs		한화손해보험	롯데손해보험	계주는행	한국화재	DB손해보험	기업은행	기업은행	한화손해보험	기업은행						
			Weights		0.23729024	0.2051397	0.18970707	0.18429979	0.18356321										
			Many to Many Pair		한화손해보험	롯데손해보험	계주는행	한국화재	DB손해보험	기업은행	기업은행	한화손해보험	기업은행						
	Cluster5	0%	Pairs		한화손해보험	롯데손해보험	계주는행	한국화재	DB손해보험	기업은행	기업은행	한화손해보험	기업은행						
			Weights		0.17095944	0.14998754	0.12696751	0.11961181	0.11035774	0.10428437	0.07896537	0.07556845	0.06329777						
			Many to Many Pair		한화손해보험	롯데손해보험	계주는행	한국화재	DB손해보험	기업은행	기업은행	한화손해보험	기업은행						
			Pairs		한화손해보험	롯데손해보험	계주는행	한국화재	DB손해보험	기업은행	기업은행	한화손해보험	기업은행						
	Cluster6	0.34%	Weights		0.17095944	0.14998754	0.12696751	0.11961181	0.11035774	0.10428437	0.07896537	0.07556845	0.06329777						
			Many to Many Pair		한화손해보험	롯데손해보험	계주는행	한국화재	DB손해보험	기업은행	기업은행	한화손해보험	기업은행						
Pairs			한화손해보험	롯데손해보험	계주는행	한국화재	DB손해보험	기업은행	기업은행	한화손해보험	기업은행								
Weights			0.17095944	0.14998754	0.12696751	0.11961181	0.11035774	0.10428437	0.07896537	0.07556845	0.06329777								

traindata: 1999.05.14~2016.01.04, testdata: 2016.01.05~2020.03.17

Window3	Cluster	수익률	traindata: 2001-06-14~2018.02.02, testdata: 2018.02.05~2022.04.13										
	Cluster1	14.13%	Many to Many Pair		한화손해보험		진양산업		진양폴리				
흥국화재			DB손해보험		진양화학								
한화손해보험			흥국화재		진양폴리		DB손해보험						
흥국화재			진양산업		진양화학		진양산업						
Weights			0.43165271		0.20626875		0.19826339		0.16381515				
Cluster2	0.11%	Many to Many Pair		BYC		성창기업지주							
		발틸		민호제강		대한방직		깨끗한나라		대한화섬			
		성창기업지주		성창기업지주		대한화섬		BYC					
		방틸		대한방직		성창기업지주		대한방직					
		Weights		0.34026218		0.27956416		0.21049861		0.16967505			
Cluster3	0.95%	Many to Many Pair		사조산업		사조오양		동원산업		신라교역			
		한성기업		CJ씨푸드		사조대림		동원수산					
		사조오양		사조오양		CJ씨푸드		사조산업		사조대림			
		한성기업		동원수산		사조오양		CJ씨푸드		사조오양			
		Weights		0.24440963		0.22293432		0.21999735		0.1867731		0.1258856	
traindata: 2001-06-14~2018.02.02, testdata: 2018.02.05~2022.04.13													
Window4	Cluster	수익률											
	Cluster1	1.37%	Many to Many Pair		사조산업		사조오양						
			한성기업		CJ씨푸드		사조대림		동원수산				
			사조오양		사조오양		사조대림		사조산업		CJ씨푸드		
			한성기업		동원수산		사조오양		CJ씨푸드		사조오양		
Weights		0.26016402		0.23151806		0.18158582		0.16548878		0.16124332			
traindata: 2002-07-02~2019.02.25, testdata: 2019.02.26~2023.04.24													
동일가중 포트폴리오 구간수익률 1.37%													

Windows	Cluster	수익률							
	Cluster1	1.43%	Many to Many Pair	삼성화재	신한지주	KT			
				기업은행	코리안리	강원랜드	한국가스공사		
			Pairs	신한지주	삼성화재	신한지주	기업은행		
				기업은행	강원랜드	코리안리	KT		
	Weights	0.31475473	0.24536827	0.24271835	0.19715866				
	Many to Many Pair	부국철강	한일철강	동양철관	NI스틸				
		DSR제강	하이스틸	휴스틸	문배철강				
	Pairs	동양철관	NI스틸	휴스틸	동양철관	한일철강	부국철강		
		하이스틸	DSR제강	동양철관	DSR제강	문배철강	DSR제강		
Weights	0.24755374	0.22359562	0.14842869	0.13256564	0.12532885	0.12252745			
Cluster2	2.75%	Many to Many Pair	사조산업	사조오양					
			한성기업	CJ씨푸드	사조대림	동원수산			
		Pairs	사조오양	사조오양	사조대림	CJ씨푸드			
			한성기업	동원수산	사조오양	사조오양			
Weights	0.33305658	0.28783763	0.2140389	0.16506689					
Cluster3	1.12%	Many to Many Pair							
		Pairs							
Weights									
동일가중 포트폴리오 구간수익률			1.77%			traindata: 2003-07-14~2020.03.09, testdata: 2020.03.10~2024.05.13			

VI. Reference

- [1] Simão Moraes Sarmento, and Nuno Horta. "Proposed Pairs Selection Framework" A Machine Learning based on Pairs Trading Investment Strategy (2020): 21-35
- [2] Hubel, David H., and Torsten N. Wiesel. "Receptive fields, binocular interaction and functional Architecture in the cat ' s visual cortex. " The Journal of physiology 160.1 (1962): 106.
- [3] Krauss, Christopher (2015) : Statistical arbitrage pairs trading strategies: Review and outlook, IWQW Discussion Papers, No. 09/2015, Friedrich-Alexander-Universität Erlangen-Nürnberg, Institut für Wirtschaftspolitik und Quantitative Wirtschaftsforschung (IWQW), Nürnberg
- [4] Chan, E. P. (2013). *Algorithmic trading: Winning strategies and their rationale*. John Wiley & Sons.
- [5] Herlemont, D. (2004). Pairs trading, convergence trading, cointegration.
- [6] Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2), 251-276.
- [7] Wang, Y., Li, X., Wu, P., Xie, H. (2023). Many-to-Many Pair Trading. In: Li, B., Yue, L., Tao, C., Han, X., Calvanese, D., Amagasa, T. (eds) Web and Big Data. APWeb-WAIM 2022. Lecture Notes in Computer Science, vol 13421. Springer, Cham. https://doi.org/10.1007/978-3-031-25158-0_31
- [8] Harlacher M (2016) Cointegration based algorithmic pairs trading. PhD thesis, University of st.Gallen
- [9] Moraes Sarmento, S., Horta, N. (2021). Proposed Pairs Selection Framework. In: A Machine Learning based Pairs Trading Investment Strategy. SpringerBriefs in Applied Sciences and Technology(). Springer, Cham. https://doi.org/10.1007/978-3-030-47251-1_3
- [10] Berkhin P (2006) A survey of clustering data mining techniques. In: Grouping multidimensional data. Springer, pp 25–71
- [11] Nonlinear Tensor Completion Using Domain Knowledge: An Application in Analysts' Earnings Forecast - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Rolling-window-with-w-28-and-n-8_fig3_349422004 [accessed 25 May, 2024]