Journal of the Korean Data Analysis Society (February 2022), 24(1), 67-81.
https://doi.org/10.37727/jkdas.2022.24.1.67

행렬 프로파일을 이용한 금융 시계열 분석*

조예나[1], 백창룡[2]

요 약

행렬 프로파일은 주어진 쿼리에서 가장 가까운 이웃을 찾는 방법인 전체 유사성 검색을 거짓 음성 없이 찾는 방법으로 Yeh et al.(2016)가 제안하였다. 본 논문은 행렬 프로파일을 금융 시계열 자료에 적용해 봄으로써 전체 유사성 검색에 기반한 자료 분석이 어떻게 활용될 수 있는지 살펴보았다. 본 논문은 2019년부터 2021년까지 관찰된 KOSPI 지수 및 실현변동성을 고려하였다. 구체적으로 행렬 프로파일을 통해 얻어진 가장 대표적인 패턴인 모티프를 통해 반복되는 추이를 살펴보았고, 가장 상이한 패턴인 디스코드를 통해 이상점을 파악하였다. 이는 탐색적 자료분석 도구로써 주식 종가에서 나타나는 대표적인 상승세와 하락세와 같은 의미 있는 정보를 제공해 주었다. 또한 유사도를 정량화한 측도를 통해서 시계열 자료의 부분 유사도를 파악하고 변곡점을 살펴보았다. 이를 변화점을 찾는 가장 표준적인 방법인 CUSUM을 이용한 결과와 비교해보았다. CAC를 이용한 결과 코로나 위기의 시작 및 백신의 개발로 인한 극복 과정에 대한 시기를 유의미하게 변화점으로 잘 탐지해내었다. 마지막으로, 행렬 프로파일을 이용한 시계열 자료의 군집화를 통해 세계 주가지수와 실현변동성들의 공동화를 살펴볼 수 있었다.

주요용어 : 행렬 프로파일, 모티프, 실현변동성, 계층적 군집화.

1. 서론

전체 유사성 검색(all-pairs-similarity-search)이란 주어진 쿼리에서 가장 가까운 이웃을 찾는 방법이다. 텍스트 도메인에서는 유사성 결합(similarity join)문제를 해결하기 위해 개발된 알고리즘들이 커뮤니티 검색, 중복 탐지, 협업 필터링, 클러스터링 및 쿼리 개선과 같은 다양한 작업에 적용되었다(Agrawal, Faloutsos, Swami, 1993). 다만 시계열에는 실행에 이르기까지의 과정이 복잡하여 전체 유사성 검색을 적용시킨 사례들은 아주 드물다. 데이터의 규모가 커질수록 두 개 이상의 시계열 간 거리 탐색의 문제는 탐색시간이 기하급수적으로 늘어 실행하기 더욱 어려워지기 때문이기도 하다. 이에 Yeh et al.(2016)은 긴 길이의 시계열도 짧은 길이의 쿼리로 잘라 한 점씩 이동하여 기존에 셀프 조인(self join)에서 발생되었던 소요 시간을 획기적으로 줄이고 활용도가 높은 방법론으로써 행

*이 논문은 제1저자 조예나의 석사학위논문의 축약본입니다.
*이 논문은 한국연구재단의 지원을 받아 수행된 기초연구 사업임(NRF-2019R1F1A1057104).
[1]03063 대한민국 서울특별시 성균관로 25-2 성균관대학교 통계학과 석사과정, 핀테크 융합전공.
E-mail : haileyohc@g.skku.edu
[2](교신저자) 03063 대한민국 서울특별시 성균관로 25-2 성균관대학교 경제대학 통계학과 교수.
E-mail : crbaek@skku.edu
[접수 2021년 12월 8일; 수정 2021년 12월 28일; 게재확정 2021년 12월 31일]

렬 프로파일을 제안하였다. 계산 시간과 관련된 이점에 대해서는 Zhu et al.(2017) 및 인용된 참조 문헌을 살펴보길 바란다. 또한 행렬 프로파일은 쿼리간 거리를 측정할 때 풀 조인(full join)을 이용하기에 거짓 음성(false negative)이 발생할 확률이 없다(Yeh et al., 2016).

행렬 프로파일을 통해 얻을 수 있는 가장 대표적인 정보는 모티프(motif)와 디스코드(discord)다. 모티프란 시계열 내에서 주기적으로 반복되는 패턴으로, 해당 시계열의 대표적인 표현, 즉 자료의 압축이 가능하다. 두 시계열을 짝지어 행렬 프로파일화 한다면 일차원적인 패턴 감지와 더불어 두 시계열간 유사성 또한 파악할 수 있다(Silva et al., 2016). 반면, 디스코드란 모티프완 반대되는 개념으로 해당 행렬 프로파일에서 다른 쿼리들과의 거리값이 비교적 큰 쿼리를 의미한다. 이는 곧 나머지 쿼리들과 가장 상이한 양상을 뛰는 것을 의미하므로 이상점 탐지(anomaly detection)에 유용하게 사용할 수 있다.

행렬프로파일이 제안된 이후에 공학, 음향학, 음성학 등 많은 시계열 자료에 대해서 경험적 분석을 하였고 그 유용성을 추가 연구를 통해서 발표하였다(Shi et al., 2019). 하지만, 금융 시계열 자료에 대한 활용은 미미하여 본 논문은 행렬 프로파일을 사용하여 금융 시계열 자료를 분석하고 이에 대한 특징을 살펴보고자 한다.

구체적으로 본 연구는 KOSPI 주가 지수 데이터와 실현변동성(realized volatility; RV) 데이터를 행렬 프로파일을 사용하여 분석하였을 때, 탐색적 자료 분석의 관점에서 시계열 자료의 어떤 특징들을 파악할 수 있는지에 대해서 중점적으로 연구하였다. 이와 더불어 유사도 분석을 통해 KOSPI와 세계 주식 시장이 어떻게 움직이는지 살펴보았다.

본 논문의 구성은 다음과 같다. 제2절은 행렬 프로파일의 생성 매커니즘과 모티프, 디스코드에 대한 정의를 기술한다. 제3절은 KOSPI 주가 지수 및 실현 변동성에 대한 실증 자료 분석 결과를 제공한다. 모티프와 디스코드와 더불어, 시계열의 부분 유사도 분석 결과를 논의하였다. 추가적으로, 행렬 프로파일을 이용해 세계 주가지수들의 군집화 양상을 확인하였다. 제4절에서는 본 논문의 결론에 대해서 요약하고 추가 논의를 기술하였다.

2. 행렬 프로파일 소개

본 장에서는 행렬 프로파일을 정의하고, 이를 통해서 계산할 수 있는 부수적인 방법론에 대해서 소개하고자 한다. 먼저 몇 가지 수식과 기호를 설명하고자 한다. 시계열 T의 부분열 $T_{i,m}$이란 i번째 위치에서 시작하여 길이 m만큼의 값들을 지닌 연속형 부분열

$$T_{i,m} = t_i, t_{i+1}, \ldots, t_{i-m+1},\ 1 \le i \le n-m+1$$

이다. 시계열 T의 모든 부분열 집합(all-subsequences set) A란 T 중 길이가 m인 윈도우(window)를 움직여 (sliding) 얻어낸 T의 모든 가능한 연속물의 순서 집합이다.

$$A = [\, T_{1,m}, T_{2,m}, \ldots, T_{n-m+1,m}].$$

이때 m은 사용자가 정의하는 부분열 길이이며, A_i는 $T_{i,m}$을 가리킨다. 시계열 T의 거리 프로파일(distance Profile) D_i는 모든 부분열 집합에 존재하는 랜덤 부분열과 기준 부분열 $T_{i,m}$간의 유클리드 거리(Euclidean distance)의 벡터이다. 즉 $d_{i,j}$가 $T_{i,m}$과 $T_{j,m}$간의 거리일 때 $D_i, (1 \le i \le n-m+1)$는 다음과 같이 정의된다.

$$D_i = [d_{i,1}, d_{i,2}, \cdots, d_{i,n-m+1}].$$

행렬 프로파일을 계산하기 위해서 전체 길이가 T인 시계열에서 길이가 m인 쿼리를 생각하자. 길이가 m인 윈도우를 한 점씩 이동하며 길이가 m인 $|T|-m+1$개의 부분열을 얻어낸다. 이 모든 부분열에 대해서 짝간 유클리드 거리를 얻어 다음과 같은 거리행렬을 만들 수 있다.

	D_1	D_2	...	D_{n-m+1}
D_1	$d_{1,1}$	$d_{1,2}$	...	$d_{1,n-m+1}$
D_2	$d_{2,1}$	$d_{2,2}$	...	$d_{2,n-m+1}$
...	...	...	...	...
D_{n-m+1}	$d_{n-m+1,1}$	$d_{n-m+1,2}$	...	$d_{n-m+1,n-m+1}$
	↓	↓	↓	↓
P	$min(D_1)$	$min(D_2)$	...	$min(D_{n-m+1})$

Figure 1. The relationship between distance matrix D and matrix profile P

이때 거리 행렬은 대칭이고 자기 자신과의 거리는 0 이므로, 대각행렬의 원소는 모두 0으로 구성된다. 이 때 얻어낸 $(|T|-m+1)\times(|T|-m+1)$ 행렬의 거리 프로파일 D_i에 대한 행렬 프로파일 P는 각 열에 대해서 가장 작은 값을 택한 벡터로 아래와 같은 식으로 표현된다.

$$P = [\min(D_1), \min(D_2), \ldots, \min(D_{n-m+1})].$$

따라서 행렬 프로파일이란 각 부분열의 가장 가까운 이웃 간의 거리가 저장되어 있는 벡터이다. 즉 첫 번째 원소값은 (자기 자신을 제외한) D_1 부분열과 가장 가까운 부분열과의 거리를 나타낸다. 또한 행렬 프로파일 인덱스 벡터란, 각 열에 대해서 최소값이 되는 부분열에 대한 위치를 저장하는 벡터다.

행렬 프로파일로 얻을 수 있는 모티프와 디스코드는 다음과 같이 계산된다. 시계열의 모티프는 시계열 내에 가장 유사한 부분열 쌍을 뜻한다. 따라서 최상위 모티프는 행렬 프로파일에서 가장 작은 값을 가지는 쌍을 나타내며, 이는 해당 부분열이 적어도 두 번 이상 비슷한 패턴으로 나타남을 의미한다. 추가적으로 최상위 모티프 부분열에 해당하는 두 점이 d_1만큼 떨어져 있다면, 모수 R을 이용해 $d_1 \times R$ 이하의 거리 값을 가지는 점들을 해당 모티프의 또 다른 짝으로 간주하게 된다. 여기서 R은 반경(radius)를 표현하는 모수이다. 하지만, 바로 이웃한 두 부분열 사이의 거리는 겹치는 부분이 많아 필연적으로 작아질 수밖에 없기에 배제구역(exclusion zone)을 설정하여 너무 인접한 부분열 짝을 찾는 트리비얼 매치(trivial match)를 피한다. 보통 원 쿼리 길이 m의 앞뒤 1/2~1배 만큼은 이웃으로 간주되지 않게 한다.

차상위 모티프짝은 최상위 모티프에서 얻어진 짝과 이웃은 제외한 후 남아있는 시퀀스 중 가장 가까운 거리에 있는 짝을 찾는다. 또한 반경 모수 R을 사용하여 차상위 모티프와 이웃한 짝들을 찾는다. 그 다음 모티프짝은 반복적으로 최상위, 차상위 모티프짝과 이웃을 제외한 시퀀스에서 가장 가까운 거리에 있는 짝을 순차적으로 찾는다.

반면 시계열의 디스코드는 행렬 프로파일 P의 가장 큰 값을 일컫는다. 따라서 거리 프로파일에서 열별 최솟값을 계산해 얻어지는 행렬 프로파일에서 가장 큰 값으로 표현된다는 것은 유사한 부분열이 없거나 매우 다른 모양새를 가진다는 것과 일맥상통한다. 그렇기에 이는 이상점 탐지기(anomaly detector)로써 매우 유용하게 사용될 수 있다.

행렬 프로파일은 또한 의미 분할(semantic segmentation)에도 활용될 수 있다. 시계열에서의 의미 분할이란 비슷한 패턴끼리의 일종의 군집화로 이해할 수 있다. 특히 Gharghabi et al.(2017) 이 제안한 CAC(Corrected Arc Crossings; CAC)를 이용하면 행렬 프로파일을 이용하여 손쉽게 시계열을 동질적인 영역인지 파악할 수 있다. 행렬 프로파일의 i번째 인덱스 벡터에는 i번째 부분열과 가장 유사한 부분열의 시작점, j 가 담겨 있다. 이 둘을 아래 Figure 2와 같이 호(arc)로 연결을 하면 각 모든 시점에 대해서 가장 가까운 부분열을 짝지을 수 있다. 예들 들어 1892번째 부분열에 가장 가까운 이웃은 1270번째 부분열이다. 각 시점에서 교차하는 호의 개수를 셀 수 있고, 많은 호가 지나간다는 것은 서로 비슷한 부분열이란 뜻이 된다. 따라서 이를 정량화함으로써 동질한 영역인지 파알할 수 있다. 보다 구체적으로 i번째 위치에서의 CAC 값은

$$CAC_i = \min\left(\frac{AC_i}{IAC_i}, 1\right)$$

으로 주어진다. IAC(idealized Arc Curve)는 시계열에 비슷한 패턴이 전혀 없는 랜덤함을 가정할 때 이론적으로 산출되는 교차하는 호의 개수로 경계효과(boundary effect)를 보정한 값이다. 따라서 CAC 값이 1에 가까우면 다른 곳과 비슷한 패턴을 보이는 것으로 이해할 수 있으며 작은 값을 가질 경우 해당 부분열과 비슷한 다른 부분열의 개수가 적기에 변곡점(changepoint)이나 특이점으로 이해할 수 있다.

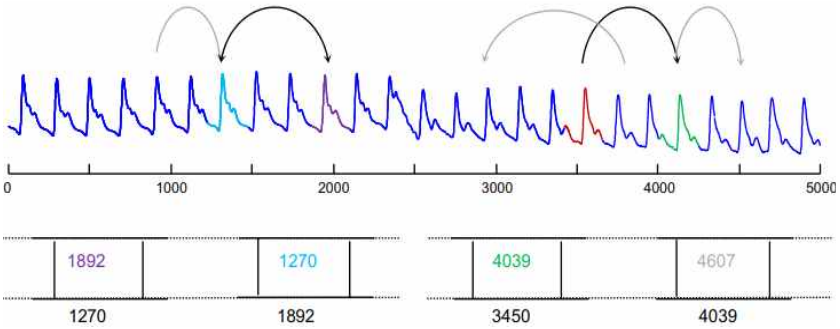

Figure 2. Selected arcs illustrated with the corresponding matrix profile indices from Gharghabi, et al.(2017)

행렬 프로파일 및 CAC에 대한 계산은 Bischoff, Rodrigues(2019)의 tsmp R-package에서 구현이 되어 있어 본 논문에서는 tsmp 패키지를 사용하여 분석을 실시하였다.

3. 실증 자료 분석

본 절에서는 행렬 프로파일을 금융 자료에 적용했을 때 탐색적 자료분석의 관점에서 어떤 정보를 얻을 수 있는지 살펴본다. 모티프를 통해 가장 대표적인 반복되는 패턴을 살펴보고자하며 디스코드를 이용해 이상점, 즉 특정 영업일에 발생하는 비정상적인 흐름을 잡아내는지 확인하고자 한다. 또한 CAC를 이용한 시계열의 유사도 분석을 통해 변곡점을 탐지하고자 한다. 마지막으로 행렬 프로파일이 부분열에 대한 거리 정보를 축약한 정보를 담고 있으므로 이를 이용하여 군집화를 진행하고 그 의미를 살펴보고자 한다.

3.1. KOSPI 종가 데이터

첫번째로 사용하는 수정된 일별 종가 데이터는 Ryan et al.(2020)의 Quantmod R-package를 이용하여 추출하였다. 분석에 쓰인 핵심 자료는 Yahoo Finance가 제공하는 2019년 1월 2일부터 2021년 6월 30일까지의 616일 동안의 수정된 KOSPI 종가 지수이다. 최근 세계 통합화로 인한 정보의 전이효과(spillover effect)를 고려해 경제 위기가 주가 변동성에 영향을 미치는 주요 지표로 강조되고 있다. 따라서 코로나와 같은 큰 경제위기가 발생한 이후 주가 양상의 변화를 확인해보고자 한다(Kim, Choi, Yoon, 2021). 또한 세계 주가 지수와의 비교를 위해서 AEX, S&P500, DAX, ALL ORDINARIES, Dow Jones, FTSE100, NIKKEI225, RUSSELL 2000, NASDAQ 100의 데이터들을 활용하였다.

3.1.1. 행렬 프로파일을 통해 발견된 모티프

Figure 3은 2019년 1월부터 2021년 6월까지의 KOSPI의 종가 데이터를 이용한 행렬 프로파일의 결과이다. 쿼리의 단위길이 m은 20으로 지정하였는데 이는 영업일 기준 한 달간의 변화를 살펴보기 위함이다. 또한 배제 구역(exclusion zone)은 쿼리와 같이 20으로 정해서 앞 뒤 한 달 동안은 이웃으로 취급하지 않는다. 반경 모수 R은 3으로 설정하였다.

Figure 3의 맨 위 패널은 KOSPI 종가의 양상에 2개의 최상위 모티프와 각각의 이웃들, 그리고 최상위 디스코드를 실선과 점선으로 표기해놓은 그래프이다. 검은색은 최상위 모티프, 빨간색은 차상위 모티프, 파란색은 최상위 디스코드를 나타낸다. 행렬 프로파일을 이용해 얻어낸 2개의 최상위 모티프는 증가하는 추세 하나와, 완만하게 감소하다가 10여일 중 약간의 반동 후 다시금 감소하는 추세로 아래 패널에 따로 표기하였다. 또한 변동폭이 점차 증가하는 20일간의 종가양상이 최상위 디스코드로 감지되었다.

Figure 3의 가운데 패널은 행렬 프로파일 P를 그린 것으로 2개의 최상위 모티프와 디스코드를 색깔로 표기한 것이다. 상단 그래프의 표기와 동일하게 최상위와 차상위 모티프 짝들을 굵은 검은 선과 붉은 선으로 표시하였고 데이터 내에서 발견된 디스코드는 파란 선으로 표기하였다. 이를 통해 행렬 프로파일에서 발견된 모티프는 최소값, 디스코드는 최대값에서 발생함을 알 수 있다.

Table 1은 앞서 기술한 옵션을 사용하여 행렬 프로파일의 모티프를 얻었을 때 발견된 최상위, 차상위 모티프와 그들의 이웃, 해당 행렬 프로파일의 값들을 표기한 것이다. 2019년 1월 2일부터 20여일간 상승하는 주가의 모양새의 모티프는 다음과 같은 발표와 맥락을 함께한다. 2019년 1월 한국은행의 금융시장 동향 보고서에 따르면 "코스피는 미.중 무역 협상을 통한 진전을 기대하며, 연준의 금리인상 속도조절을 시사하고, 중국 정부의 경기 부양책 발표 등으로 투자심리가 개선되면서 상승" 한다고 밝혔다. 18년 12월 말에 2,041이었던 코스피는 19년 1월 말에 2,205 , 연이은 2월 14일은 2,226으로 계속해서 상승하였다. 또한 "금년 1월 이후 국제금융시장은 미 연준의 금리인상 속도조절 시사 이후 투자심리가 크게 회복" 하고 있다고 주목했다.

차상위 모티프중 먼저 발생하는 2019년 4월 15일부터 20일간의 종가의 하락세는 동일한 한국은행의 금융시장 동향 보고서에 의거하여, "코스피는 4월 들어 오름세를 지속하다가 중순 이후 IT업종 회복 지연 우려로 미·중 무역협상 불확실성 증대 등으로 하락"하였다고 언급된다. 구체적으로는 4월 16일 2,249에서 4월 말 2,204 , 5월 10일 2,108로 전반적인 하락하는 모양새를 보였다. 이는 "중순 이후 IT 업종 회복 지연의 우려와 미.중 무역협상의 불확실성 증대로 인한 하락"이라고 밝혔다.

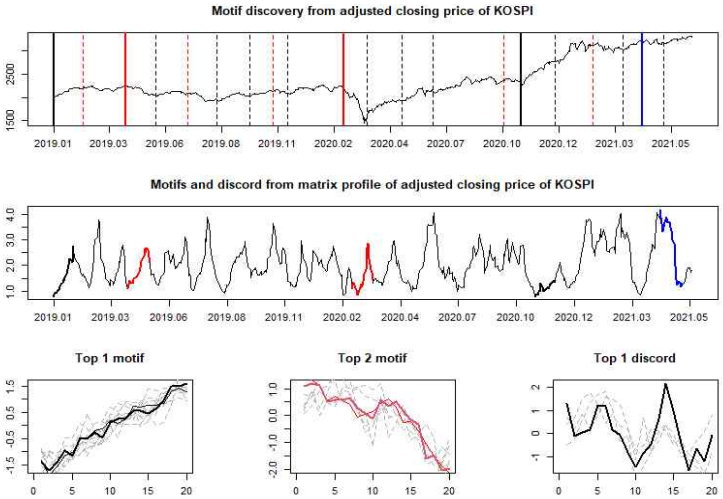

Figure 3. top) Motif and discord discovery from adjusted closing price of KOSPI middle) Motifs and discord discovered in matrix profile of adjusted closing price of KOSPI bottom) Top 2 motifs and top 1 discord from adjusted closing price of KOSPI

최상위 디스코드는 2021년 4월 16일부터 한달 동안 나타난 패턴으로 주가가 갑작스럽게 하락하며 그를 유지하다 반등하는 매우 불안정한 금융시장의 패턴을 보여준다. 이 시점에 한국은행 동향 보고서에 따르면, “국내외 경기회복을 기대하는 여론과 주요 기업의 양호한 실적 발표 등으로 상당폭 상승”하였다 보도했음을 미루어 보아 코로나 위기로 인한 불안감과 경기회복에 대한 기대가 공존하는 혼동스러운 모습을 디스코드가 찾아낸 것을 알 수 있다.

Table 1. Discovered motifs, discord and their neighbors from adjusted closing price of KOSPI

Start date	End date	Type	Matrix profile
2019-01-02	2019-01-30	motif 1	0.789
2020-10-29	2020-11-26	motif 1	0.789
2019-05-27	2019-06-25	neighbor of motif 1	1.2499
2019-08-23	2019-09-24	neighbor of motif 1	1.0099
2020-03-24	2020-04-22	neighbor of motif 1	1.3892
2020-05-12	2020-06-09	neighbor of motif 1	1.0509
2021-03-23	2021-04-20	neighbor of motif 1	0.8564
2021-05-21	2021-06-03	neighbor of motif 1	1.2878
2019-04-15	2019-05-15	motif 2	1.1054
2020-02-19	2020-03-18	motif 2	1.1054
2019-02-13	2019-03-14	neighbor of motif 2	1.5478
2019-07-12	2019-08-09	neighbor of motif 2	1.0664
2019-10-22	2019-11-20	neighbor of motif 2	1.7911
2019-11-08	2019-12-09	neighbor of motif 2	2.1231
2019-12-02	2020-01-02	neighbor of motif 2	1.1506
2020-06-26	2020-07-24	neighbor of motif 2	1.4825
2020-10-05	2020-11-03	neighbor of motif 2	2.1231
2020-12-16	2021-01-18	neighbor of motif 2	1.1531
2021-02-09	2021-03-12	neighbor of motif 2	2.6213
2021-04-16	2021-05-17	discord	4.1479

3.1.2. CAC를 이용한 유사도 분석

Figure 4는 행렬프로파일을 이용해 유사도 분석을 진행한 그림을 보여준다. 먼저 맨위 패널은 CAC 값을 나타낸다. 2020년 3월 16일 그리고 2020년 12월 22일에 CAC그래프가 크게 변화함을 한눈에 알 수 있다. 즉 이 두 시점 근처의 부분열과 유사한 패턴을 보이는 다른 부분열을 전체 시계열에서 찾을 수 없음을 의미하여 이 두 점이 바로 변곡점에 가깝다고 생각할 수 있을 것이다. 이 두 변곡점을 KOSPI 주가와 함께 중간 패널에 표시하였다. 또한 기존의 변화점을 찾는 가장 표준적인 방법인 CUSUM을 이용하였을 때 찾은 세 변화점을 마지막 패널에 나타냈다. CUSUM을 이용해 발견된 변화점은 2020년 2월 27일, 2020년 7월 10일, 2020년 12월 8일이다. CUSUM의 변화점 개수는 Zeileis et al.(2015)의 strucchange R-package에서 제안한 BIC(Bayesian Information Criteria)를 이용하였으며 CUSUM을 이용한 변화점 탐구는 Lee, Baek(2020) 및 인용된 참조문헌들을 참조하기 바란다. 각 변화점 수에 따른 BIC는 Table 2와 같다. 따라서 m이 3 이상일 때의 BIC의 변화가 미미하므로 최적의 구분점 수는 3으로 간주하였다.

Table 2. The number of segmentation from adjusted closing price of KOSPI using CUSUM and its correspoding BIC

m	0	1	2	3	4	5
BIC	9256	8104	7795	7717	7710	7681

놀랍게도 CAC를 이용하여 찾은 두 변화점이 CUSUM을 이용해서 찾은 변화점과 매우 유사함을 찾을 수 있다. 또한 2020년 3월 19일은 코로나 위기로 인해 KOSPI 지수가 연중 최저를 기록한 날이다. 2020년 11월에는 코로나 백신이 개발되었고 2020년 12월 3일에는 영국에서 최초로 코로나 백신 접종이 시작되었다. 즉 3월 16일을 경계로 그 이전은 코로나 위기 이전의 경제 상황을 하나의 유사한 군집으로 찾고, 두 번째 변화점인 2020년 12월 22일까지 두 번째 구간에서는 코로나 위기 이후 이에 대한 극복의 과정이라 묶어서 생각할 수 있으며 2020년 12월 22일 이후에는 코로나 사태가 장기화 되는 과정 중 백신의 발표로 인해 주식 시장 상황이 다른 국면으로 전개되면서 위기를 극복해 나가는 과정을 보여준다고 해석이 가능하다. 즉 코로나 위기 이전, 코로나 위기 중 및 백신 개발 이후 코로나 극복에 따라 KOSPI 시장이 각기 다른 양상으로 전개되고 있음을 CAC를 이용한 유사도 분석에서 찾을 수 있었다.

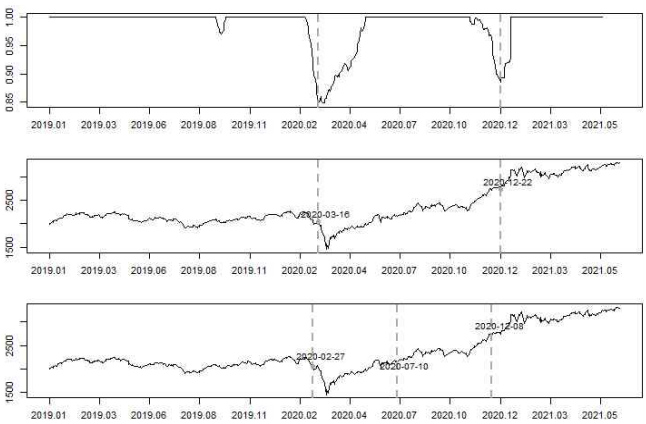

Figure 4. top) CAC plot from adjusted closing price of KOSPI middle) segmentation by FLUSS bottom) segmentation by CUSUM

3.1.3. 행렬프로파일을 이용한 군집화

본 절에서는 2019년 1월 2일부터 2021년 5월 31일까지의 KOSPI를 포함하여 대표적인 세계 주가 지수인 AEX, S&P500, DAX, ALL ORDINARIES, Dow Jones, FTSE100, NIKKEI225, RUSSELL 2000, NASDAQ 100의 주식 종가 데이터들을 각기 행렬 프로파일로 전환한 뒤 계층적 군집화를 진행 하였을 때의 분류 결과를 살펴보고자 한다. 경제 분야에서는 시계열의 패턴이 유사한 거래 종목들을 군집화하여 자료의 변동성의 여러 가지 특성을 설명할 수 있어 유용하기 때문이다(Kwon, Park, 2020). 계층적 군집화에서는 군집의 개수를 정하는 것이 매우 중요한데 본 논문은 Duda et al.(1973)의 두다 인덱스(duda index)를 이용하여 최적 군집의 수를 설정하였다. 두다 인덱스는 군집을 고려했을 때와 그렇지 않았을 때의 오차제곱합의 비율을 기준으로 군집의 수를 결정한다. 최적의 군집의 수는 Gordon(1999)이 제안한 임계값의 기준을 사용하였으며 R-package 인 NbClust(Charrad et al., 2014)를 사용하여 계산하였다. 두다 인덱스가 찾은 최적의 군집의 수는 3이며 군집화 결과를 Figure 5에 나타냈다.

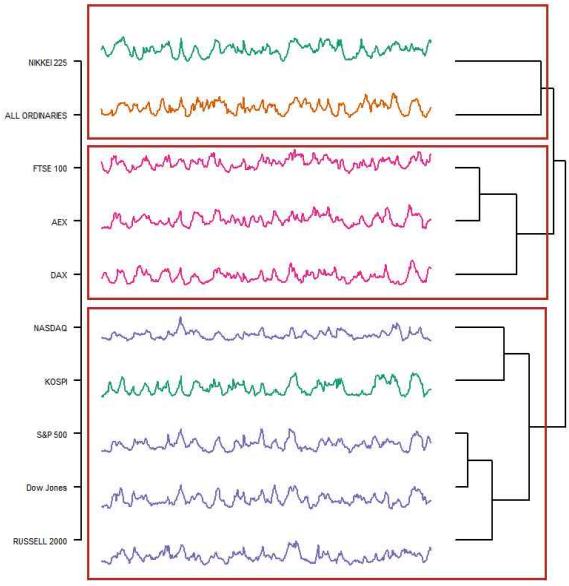

Figure 5. Hierachical clustering result from matrix profile of adjusted closing price data

사용한 전 세계 10개 주가 지수를 대륙별 다른 색을 사용하여 표현하였다. 북미주는 보라색, 유럽주는 분홍색, 아시아주는 초록색, 오세아니아주는 주황색을 사용하였다. 먼저 세 개의 군집은 일본 및 오세아니아주가 하나의 군집으로 형성되고 있고, FTSE100, AEX, DAX 유럽 지수가 다른 하나의 군집을 이룬다. 마지막으로 많은 연구에서 외환위기 이후 우리나라 주식시장은 미국과 아시아 시장과의 동조화가 심화되고 있다는 것을 받침 삼아 미국 및 한국의 KOSPI가 하나의 군집으로 결집됨을 살펴볼 수 있다(Kim, Park, Kim, 2007). 즉 한국의 주가지수 KOSPI는 유럽이나 일본 혹은 오세아니아주 보다는 미국과 비슷한 양상을 띠는 것으로 파악되어 미국 주가 지수의 영향을 가장 많이 받는 공동화(comovement)가 존재함을 알 수 있다. 이는 국제무역으로 인한 경제의 상호의존성이 높아짐에 따라 국제적인 주식거래가 활발해진 것에 따른 것이라고 해석될 수 있다(Kang, 2012).

3.2. 실현변동(Realized variance) 분석

본 논문에서 분석한 두 번째 자료는 실현변동성(Realized Volatility; RV)으로 금융시장의 변동성에 대해서 행렬 프로파일을 적용했을 때 어떠한 특징을 찾을 수 있는지 살펴보았다. 변동성(volatility)은 통상 로그-수익률(log-return)로 t시점과 일별 종가 P_t에 대하여

$$r_t = \log P_t - \log P_{t-1}$$

로 정의된다. 하지만, 고빈도 자료가 발달함에 따라 Anderson et al.(2003)은 이를 이용하여 보다 정확하게 추정할 수 있는 지표인 실현 변동성을 다음과 같이 제안하였다.

$$RV_t = \sqrt{\sum_{j=0}^{M-1} r_{t-j\cdot\Delta}^2}\,.$$

즉 적절한 빈도 Δ에서 얻어낸 하루 중(intraday)의 수익률, r_t의 거듭 제곱의 합으로 추정하였다. 본 논문에서는 5분 간격으로 관측한 주가 지수에 대해서 옥스포드만의 양적 금융 실현 라이브러리(https://realized.oxford-man.ox.ac.uk)에서 제공한 자료를 분석하였다. 2019년 1월 2일부터 2021년 5월 30일까지 자료이며 AEX, AEX, S&P500, DAX, KOSPI, ALL ORDINARIES, Dow Jones, FTSE100, NIKKEI225, RUSSELL 2000, NASDAQ 100 주가지수의 변동성을 분석하였다.

3.2.1. 행렬 프로파일을 통해 발견된 모티프와 디스코드

Figure 6은 2019년 1월부터 2021년 5월까지 KOSPI 종가의 실현변동성 데이터를 이용하여 얻어낸 디스코드와 모티프를 기술한 그림이다. 이 분석에서 사용한 쿼리의 길이는 25이며, 반경 $R = 1.5$ 배제 구역은 1로 설정하였다. 쿼리 길이의 경우 앞서 종가 데이터에서 사용한 길이 20일의 경우 뚜렷하게 구별되는 모티프를 찾아내기 어려워서 5일 가량의 기간을 추가하여 변동성을 분석하였다.

상단 패널은 KOSPI 실현변동성의 시계열도표 및 행렬 프로파일을 사용하여 찾은 모티프와 그의 짝들의 위치를 표시하였다. 검정색은 최상위, 빨간색은 차상위 모티프, 파란색은 최상위 디스코드를 나타낸다. 최상위, 차상위 모티프 및 디스코드는 하단 패널에 더 자세히 그렸다. 먼저 최상위 모티프는 변동성이 하락하는 패턴이다. 이는 주가 시장이 안정을 찾아가고 있는 양상을 의미한다. 차상위 모티프는 변동성이 해소된 듯 천천히 줄어들었다가 다시 급등하는 형태로 외부 요소나 내부 조정등으로 시장의 불안이 급증하는 모양이다. 최상위 디스코드의 경우 변동의 추세 변화 없이 상승과 하락을 빈번히 반복하는 매우 불안정한 모습이 관찰되었다.

Table 3은 KOSPI의 실현변동성 데이터에서 얻은 모티프와 디스코드들의 날짜를 나타낸다. 최상위 모티프는 2019년 8월 6일부터 25일간 발생하였는데 이 시기 한국은행 금융시장 동향 보고서에 따르면 "코스피는 대외 불확실성 증대 및 국내기업 실적 부진의 영향으로 상승폭 하락하였다가 미·중 무역협상 재개, 홍콩사태 완화 등으로 큰 상승"을 보였다고 기록하고 있다. 즉 금융 시장이 호재로 인하여 불안감이 해소되고 안정을 찾아가고 있음을 확인할 수 있었다. 차상위 모티프는 2020년 1월 30일부터 2020년 3월 6일까지로 차상위 모티프는 코로나 위기로 인해서 주가가 연일 폭락하고 시장이 패닉에 빠진 상황을 잘 나타내주고 있다. 또한 최상위 디스코드는 2019년 10월 11일부터 2019년 11월 15일까지 발생하는 것으로 파란 선으로 표시하였다. 이때의 한국은행 보고

서에 따르면 “미·중 무역협상 진전 및 반도체 업황 회복 기대 등으로 상승하였다가 11월 중순 이후 미·중 무역협상 불확실성 등이 부각되면서 상당폭 하락”하는 모습을 보여 큰 변동폭을 보임을 알 수 있었다.

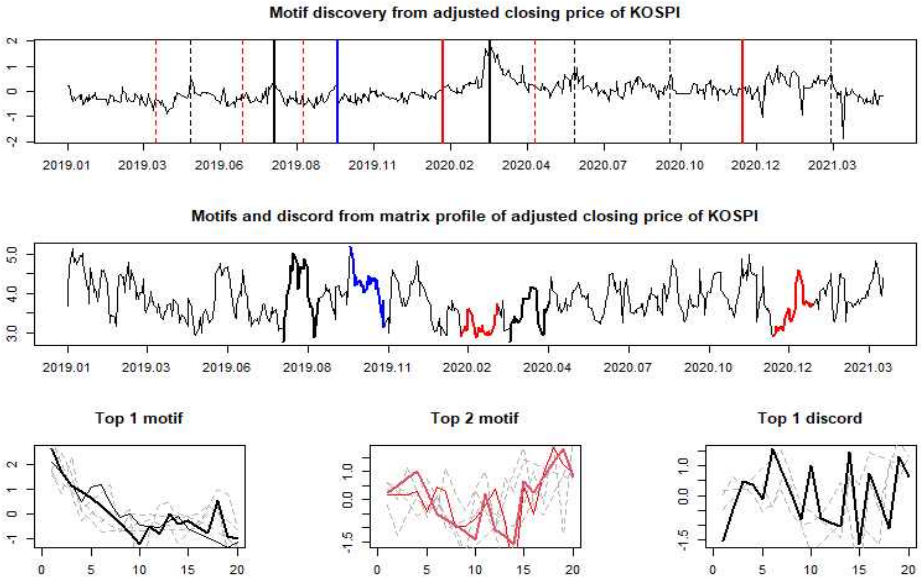

Figure 6. top)Motif and discord discovery from RV of KOSPI middle) Motifs and Discord discovered in matrix profile of RV of KOSPI bottom) Top 2 motifs and top 1 discord from RV of KOSPI

Table 3. Discovered motifs, discord and their neighbors from realized volatility of KOSPI

Start date	End date	Type	Matrix profile
2019-08-06	2019-09-10	motif 1	2.7715
2020-03-19	2020-04-27	motif 1	2.7715
2019-05-10	2019-06-14	neighbor of motif 1	2.9485
2020-06-15	2020-07-20	neighbor of motif 1	3.2675
2020-09-22	2020-10-29	neighbor of motif 1	3.6796
2021-03-09	2020-04-02	neighbor of motif 1	4.0369
2020-01-30	2020-03-06	motif 2	2.8933
2020-12-07	2021-01-04	motif 2	2.8933
2019-04-04	2019-05-14	neighbor of motif 2	3.5834
2019-07-04	2019-08-08	neighbor of motif 2	3.3037
2019-09-05	2019-10-10	neighbor of motif 2	3.1115
2020-05-04	2020-06-08	neighbor of motif 2	3.9702
2019-10-11	2019-11-15	discord	5.1899

3.2.3. CAC를 이용한 실현변동성 유사도 분석

실현변동성에 대한 유사도 분석은 Figure 7에 요약하였다. 상위 패널은 CAC값을 나타내며 2019년 7월 1일과 2020년 9월 25일 두 시점에서 뚜렷한 패턴의 변화가 일어남을 알 수 있다. 특히 두 번째 변곡점인 2020년 9월 25일은 2020년 3월부터 줄곧 감소하는 추세를 보여줌으로써 부분열에

대한 유사도가 계속 떨어지고 있음을 보여주고 있다. 이는 앞선 모티프 분석 결과와 같이 코로나로 인한 금융시장의 충격으로 인한 혼란이 가중되고 있음을 보여준다고 할 수 있다. CAC 그림은 이러한 패닉이 2020년 9월에 들어서야 진정이 되어 새로운 국면으로 전환됨을 보여준다. 실현 변동성을 CUSUM을 이용하여 변화점을 찾은 경우 CAC를 이용하여 찾은 변곡점과는 다른 결과를 줌을 알 수 있다. 맨 아래 패널에 CUSUM의 Table 4에서 제시된 바와 같이 BIC를 이용한 변화점은 역시 2개로 같았으나 날짜가 2020년 2월 18일 및 2020년 6월 25일로, 변동성이 큰 폭의 상승을 기록한 점을 변화점으로 찾았다. CAC는 부분열에 대한 유사도를 측도화한 것으로 평균이 큰폭으로 오른 지점을 찾은 CUSUM과는 미묘하게 다른 의미를 가지고 있음을 알 수 있다. 즉 CAC는 국소화(localization)된 부분열에 대한 유사도이고, CUSUM은 전체 시계열에서 평균이나 분산의 변화와 같은 값의 변화로 두 방법이 지향하는 바가 변화를 탐지하는 데에는 같지만 어떠한 지점을 변화로 인식하는 지에 대해서는 서로 다른 점이 있음을 확인할 수 있었다.

Table 4. The BIC for the selection of segmentation for KOSPI realized volatility.

m	0	1	2	3	4	5
BIC	608.39	408.73	340.25	343.58	353.60	365.04

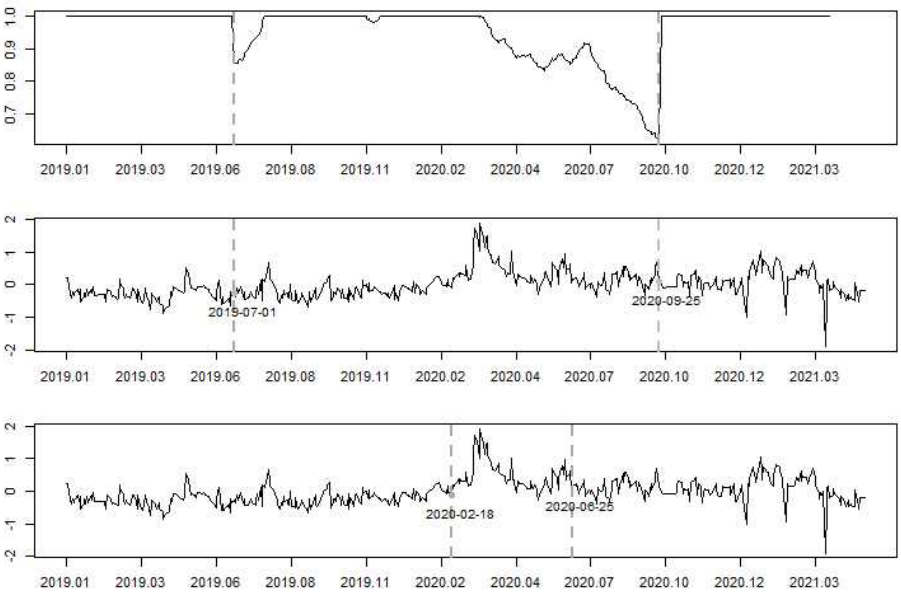

Figure 7. top) CAC plot from realized volatility of KOSPI middle) segmentation by FLUSS bottom) segmentation by CUSUM

3.2.4. 행렬프로파일을 이용한 군집화

세계의 실현변동성 자료를 통해 계산된 행렬 프로파일을 사용하여 계층적 군집화를 실시하였고 Figure 8에 그 결과를 나타냈다. 가독성을 위해 아시아, 오세아니아, 북미, 유럽주의 색을 각각 초록색, 주황색, 보라색, 분홍색으로 표기하였다. 최종 군집의 개수는 두다 인덱스를 사용하였으며 최적

의 군집의 개수는 4개 이다. 실현변동성의 경우 대륙간 군집화의 양상을 띠던 주가 지수와는 다른 결과를 찾을 수 있었다. 대표적인 미국 지수인 NASDAQ 100, S&P500, Dow Jones는 하나의 군집을 이루어 비슷한 실현변동성을 보임을 확인할 수 있었다. 또 유럽지수인 AEX와 DAX 역시 하나의 군집으로 묶여 있음을 확인하여 지역적인 특색을 찾아볼 수도 있었다. 하지만, KOSPI를 비롯 RUSSEL 1000과 NEKKEI225는 다른 지역주임에도 불구하고 하나의 군집으로 묶여있어 변동성의 경우 주가 지수와는 사뭇 다른 양상으로 전개됨을 살펴볼 수 있었다. 변동성은 주가 지수의 상승 혹은 하락의 패턴과는 다른 양상으로 전개됨을 보여준다고 할 수 있겠다. 이와 관련해서는 추가 연구를 통해 행렬 프로파일을 사용한 군집화가 금융시장의 변동성을 이해하는데 더 좋은 정보를 제공해주는지 확인해볼 필요가 있다. 예를 들어, Gharghabi et al.(2020)은 행렬 프로파일로 얻어낸 거리 측도가 현재 존재하는 유클리드 거리와 같은 다른 거리 측도들보다 데이터 외부에서 흔히 발생하는 급증(spike), 중도 탈락(dropout), 불안정한 기준선(wandering baseline), 결측값 등에 강건하다고 언급하고 있다.

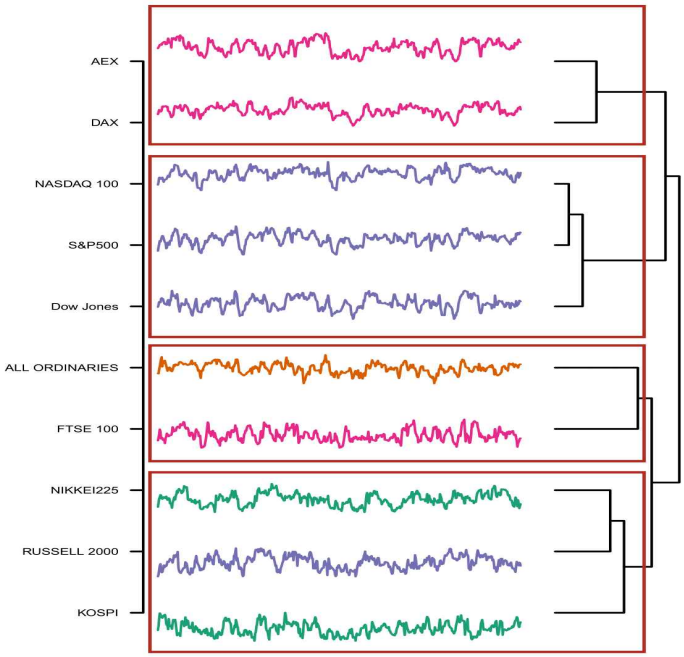

Figure 8. Hierachical clustering result from matrix profile of realized volatility

4. 결론 및 논의점

행렬 프로파일은 시계열 자료의 유사성 검색을 획기적으로 빠르게 한 방법으로 대표적인 패턴인 모티프, 가장 상이한 디스코드 및 유사도를 파악하는데 유용한 방법이다. 본 논문은 이러한 행렬 프로파일의 특징을 이용하여 금융시계열 자료의 탐색적 도구로서의 활용방안에 대해서 모색해 보았다. 좀 더 구체적으로 본 논문은 2019년 1월부터 2021년 6월까지의 KOSPI 수정 종가를 비롯 실현변동성을 분석해 보았다. 그 결과 찾은 행렬프로파일 모티프는 주가의 상승세와 하락세를 적

절하게 찾았으며 CAC 그래프의 경우 코로나 위기의 시작 및 백신의 개발로 인한 극복 과정에 대한 시기를 유의미하게 변곡점으로 찾았다. 또한 행렬프로파일을 통한 군집화는 각 대륙간 주가 지표로 군집화를 잘 시켰으며 특히 우리나라 KOSPI의 경우 미국의 시장과 공동화가 일어남을 살펴볼 수 있었다. 실현변동성에 대해서도 행렬 프로파일은 탐색적 자료분석 도구로서 의미있는 정보를 제공해 주었다. 모티프를 통한 변동성이 감소 혹은 증가하는 패턴을 찾았고 디스코드를 통해서는 불안정한 변동성 구간을 제공해주었다. 하지만 CAC의 경우 CUSUM 변화점과는 사뭇 다른 시점을 골라 행렬 프로파일이 국지화된 부분열의 변화를 탐지해 냄을 살펴볼 수 있었다. 행렬 프로파일을 통한 군집화 역시 유럽 및 미국의 시장에 대해서는 대륙간 군집화를 잘 했지만, KOSPI의 경우 일본 및 RUSSEL 2000과 군집화를 이루어 주가 지수와는 차별된 결과를 도출해주었다.

이렇듯 행렬 프로파일은 유사성에 기반한 자료 탐색도구로서 시계열 분석에 의미있는 정보를 전달해 주었다. 또한 행렬 프로파일은 일종의 차원 축소의 역할도 제공하여 본래의 시계열 자료에 대해 새로운 시각을 제공해 주기도 하였다. 하지만, 변동성 분석에서도 나타났듯이, 국지화에 기반한 분석이기에 기존의 방법론과의 비교 분석을 통해 장단점을 더 심층적으로 연구할 필요가 있다. 또한 행렬 프로파일에 쓰인 쿼리의 길이 m에 따라 분석결과가 달라져 이를 통계적으로 결정하는 방법론에 대한 개발도 추가 연구가 필요하다.

References

Agrawal, R., Faloutsos, C., Swami, A. (1993). Efficient similarity search in sequence databases, International conference on foundations of data organization and algorithms, *Springer*, 69-84. DOI: 10.1007/3-540-57301-1_5

Bischoff, F., Rodrigues, P. P. (2019). *Tsmp: An R Package for Time Series with Matrix Profile*, arXiv preprint arXiv:1904.12626. DOI: 10.32614/RJ-2020-021

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set, *Journal of Statistical Software*, 61, 1-36.

Gharghabi, S., Ding, Y., Yeh, C. C. M., Kamgar, K., Ulanova, L., Keogh, E. (2017). *Matrix Profile VIII: Domain Agnostic Online Semantic Segmentation at Superhuman Performance Levels*, 2017 IEEE international conference on data mining (ICDM), 117-126. DOI: 10.1109/ICDM.2017.21

Gharghabi, S., Imani, S., Bagnall, A., Darvishzadeh, A., Keogh, E. (2020). An ultra-fast time series distance measure to allow data mining in more complex real-world deployments, *Data Mining and Knowledge Discovery*, 34, 1104-1135. DOI: https://doi.org/10.1007/s10618-020-00695-8

Kang, I. C. (2012). An empirical study on efficient forecasting method of Korea and advanced stock market's volatility, *Journal of The Korean Data Analysis Society*, 14(4), 2125-2138. (in Korean).

Kim, B.-K., Choi, K.-H., Yoon, S.-M. (2021). Effects of macroeconomic variables, global economic uncertainty, and sentiment on korean stock market volatility: application of a GARCH-MIDAS mode, *Journal of The Korean Data Analysis Society*, 23, 1699-1715. (in Korean). DOI: https://doi.org/10.37727/jkdas.2021.23.4.1699

Kim, T. H., Park, J. H., Kim, M. R. (2007). Analysis of volatility clustering and asymmetry and volatility forecasting in KOSPI, *Journal of The Korean Data Analysis Society*, 9(6), 2861-2875. (in Korean). DOI: https://doi.org/10.37727/jkdas.2020.22.1.215

Kwon, S, Park. M. S., (2020). Time-series data clustering based on the correlation of periodogram, *Journal of the Korean Data Analysis Society*, 22(5), 1751-1766. (in Korean). DOI: https://doi.org/10.37727/jkdas.2020.22.5.1751

Lee, T., Baek, C. (2020). Block wild bootstrap-based CUSUM tests robust to high persistence and misspecification, *Computational Statistics and Data Analysis*, 150, 106996. DOI: https://doi.org/10.1016/j.csda.2020.106996

Ryan, J. A., Ulrich, J. M., Thielen, W., Teetor, P., Bronder, S., Ulrich, M. J. M. (2020). *Package 'Quantmod'*.

Shi, J., Yu, N., Keogh, E., Chen, H. K., Yamashita, K. (2019). *Discovering and Labeling Power System Events in Synchrophasor Data with Matrix Profile*, 2019 IEEE Sustainable Power and Energy Conference (iSPEC), 1827-1832. DOI: 10.1109/iSPEC48194.2019.8975286

Silva, D. F., Yeh, C. C. M., Batista, G. E., Keogh, E.. J. (2016). *SiMPle: Assessing Music Similarity using Subsequences Joins*, ISMIR, 23-29. DOI: https://doi.org/10.5281/zenodo.1415012

Yeh, C. C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Silva, D. F., Mueen, A., and Keogh, E. (2016). *Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets*, 2016 IEEE 16th international conference on data mining (ICDM), 317-1322. DOI: https://doi.org/10.1109/ICDM.2016.0179

Zeileis, A., Leisch, F., Hornik, K., Kleiber, C., Hansen, B., Merkle, E. C., Zeileis, M. A. (2015). *Package 'Strucchange'*.

Zhu, Y., Zimmerman, Z., Senobari, N., S., Yeh, C.-C. M., Funning, G., Mueen, A., Brisk, P., Keogh, E. (2016). *Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins*, 2016 IEEE 16th international conference on data mining (ICDM), 739-748. DOI: https://doi.org/10.1109/ICDM.2016.0085

Matrix Profile as an Exploratory Financial Data Analysis Tool*

Yena Cho[1], *Changryoung Baek*[2]

Abstract

Matrix profile proposed by Yeh et al. (2016) finds the nearest neighbors at a given query without false negatives, which is also called all-pair-similarity search. In this paper, we examined matrix profile as an exploratory data analysis tool for financial time series. To be more specific, we consider KOSPI index and realized volatility from 2019 to 2021. We analyzed motif to find the most representative pattern of the financial time series, and discord to identify outlying patterns. As an exploratory data analysis tool, it provided meaningful information such as a representative rise and decline in stock closing prices. Also, we conducted semantic segmentation to detect changepoints and locate similar regimes. we compared the results with CUSUM, the most standard method of finding changepoint. As a result of using CAC, the timing of the start of the COVID-19 crisis and the overcoming process due to the development of vaccines was well detected as a significant change. Finally, hierarchical clustering using matrix profile provided meaningful comovement of KOSPI market with US markets.

Keywords : matrix profile, motif, realized volatility, hierarchical clustering.

*This work was supported by the Basic Science Research Program from the National Research Foundation of Korea (NRF-2019R1F1A1057104).

[1](First Author) Graduate student, Department of Statistics, Dept. of Fintech, Graduate School, Sungkyunkwan university, Sungkyunkwanro 25-2, Seoul, 03063, Korea. E-mail : haileyohc@g.skku.edu

[2](Corresponding Author) Professor, Department of Statistics, Graduate School, Sungkyunkwan university, Sungkyunkwanro 25-2, Seoul, 03063, Korea. E-mail : crbaek@skku.edu

[Received 8 December 2021; Revised 28 December 2021; Accepted 31 December 2021]