주성분 분석을 활용한 팩터 분석

+최문석 ++박지현
경북대학교 금융데이터 분석학회 DART
cmschs0301@knu.ac.kr, blackhole124@knu.ac.kr

Factor analysis with Principal Component Analysis

Choi Mun Seok Park Ji Hyun

요 약

금융에서 다루는 요인 모형은 거시경제학적 요소, 기업의 가치를 활용하는 방법, 통계적 방법 등이 있다. 이는 외부요인을 이용하는 방법과 자산들의 수익률에 내포된 내재 요인들을 추정하는 방법이 있다. 본 고에서는 주성분 분석(PCA, Principal Component Analysis)를 사용해 KOSPI200, S&P500 기업들의 수익률에 내포된 내재 요인을 추출해 비교한다. 이러한 분석은, 숨겨져 있는 독립적인 요인을 통계적 방법을 통해 찾는 독립성분분석(IPA)의 일종으로, 요인 분석과 포트폴리오 구축 등에 있어 다양하게 활용할 수 있다.

1. 서 론

주식의 수익률과 같은 금융 시계열 자료는 복잡하고 다양한 소음이 섞여 있어 주요한 요인을 찾아내기가 어렵다. PCA를 통한 요인 추출은 혼합된 신호의 정보만을 이용해 원래의 신호를 추정해 내는 기법의 하나이다. PCA는 여러 변수로 이뤄진 데이터를 보고 이 데이터를 가장 잘 설명할 수 있는 주성분을 찾는 기법이다. 각 주성분은 서로 직교 관계를 지니므로 서로 독립이다. KOSPI200, S&P500 기업들에 대해 PCA를 적용해서 주성분을 추출한 후 이를 한국 및 미국 내 지수와 비교한다.

2. 관련 연구

[2] Factor Models, Machine Learning, and Asset Pricing, Annual Review of Financial Economics, Stefano Giglio, Bryan Kelly, Dacheng Xiu, 2022에 따르면 PCA를 통해 잠재인자와 로딩을 추출하는 연구는 다음과 같다. PCA의 사용은 Chamberlain & Rothschild (1983) 및 Connor & Korajczyk (1986)부터 시작되었으며, 최근에는 Kozak, Nagel & Santosh (2018), Kelly, Pruitt & Su (2019), Pukthuanthong, Roll & Subrahmanyam (2019), Giglio & Xiu (2021)과 같은 연구에서 더욱 인기를 얻고 있다. 최근에는 PCA의 단점을 보완한 리스크 프리어 PCA 추정 및 조건부 인자 모델과 같은 PCA 기법이 나오고 있다.

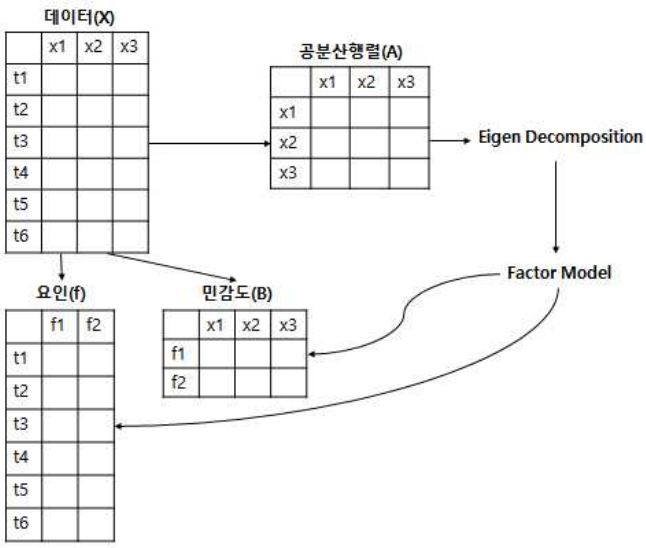
3. 이론적 배경

3.1 주성분분석(PCA, Principal Component Analysis)

주성분 분석이란 데이터의 피처를 압축하여 데이터의 차원을 낮추는 것이다. 여러 개의 요소가 갖는 정보를 하나로 압축하는 것이다. 즉, 유사한 요소들은 하나의 요소로 합쳐진다. 이때, 유사하다는 것을 판단할 때 사용하는 것이 주성분 분석이다. 공분산 행렬을 이용해 고유 벡터 및 고윳값을 구하고, 분산이 가장 큰 방향을 가진 고유벡터에 입력 데이터를 선형 변환한다. 그 후, 고유벡터와 직교하며 분산이 가장 큰 고유벡터로 선형 변환을 한다. 위와 같은 과정을 반복하여 필요한 만큼의 고유벡터를 구한다. 이때, 고유벡터는 요인 모형을 추정할 수 있는 근거가 된다.

3.2 요인 모형(Factor Model)

[1] Factor Models, MIT OpenCourseWare에 따르면, 요인 모형은 k개의 공통 요인과 민감도로 주가 수익률을 설명하는 경제 모델이다. 이 모형을 활용하기 위해서는 주가 수익률 데이터로부터 요인과 민감도를 추정해야 한다. 우리는 수익률 데이터 X에서 공분산 행렬 A를 추정한다. 그렇게 하면, 해당 행렬에 대한 고윳값과 고유벡터를 찾을 수 있다. 다음으로 고윳값과 고유 벡터를 이용해서 요인 모형을 표현한다. 고유 벡터는 민감도 B를 추정할 수 있고, 요인은 X와 B의 전치행렬 간의 행렬곱을 통해 추정할 수 있다. 그림1은 이러한 과정을 도식화한다.



<그림1 : 주성분분석을 통한 요인 분석>

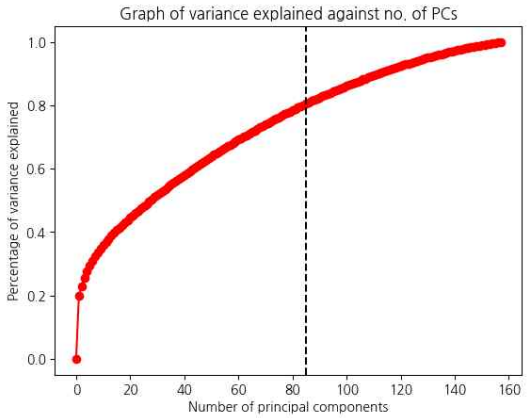
4. 분석 결과

4.1 KOSPI 200

2013년 11월 13일부터 2023년 11월 10일까지의 현재 기준 코스

+ 경북대학교 식품자원경제학과
++ 경북대학교 컴퓨터학부

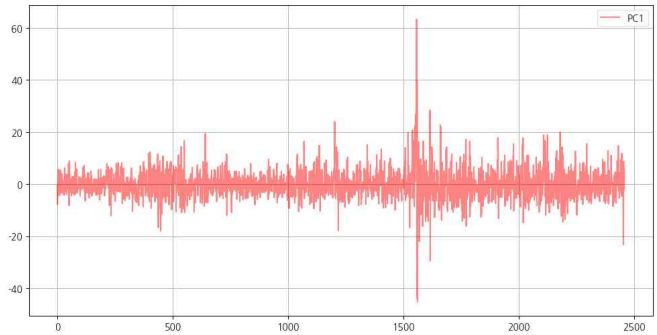
피 200종목들의 종가를 대상으로 분석을 진행하였다. 해당 기간 내의 신규 상장 종목들은 제거하고 분석을 진행하였다. 종가 데이터를 수익률 데이터로 변환 후, PCA를 위해 정규화를 진행한다. 그 후, 전체 분산의 약 80%를 설명하는 85개의 주성분(PC, Principal Component)를 선정한다. 나머지 주성분과 고윳값의 차이가 큰 앞의 3개의 주성분만 자세히 분석하기로 한다.



<그림2 : KOSPI200, 고윳값을 활용한 주성분 개수 추정>

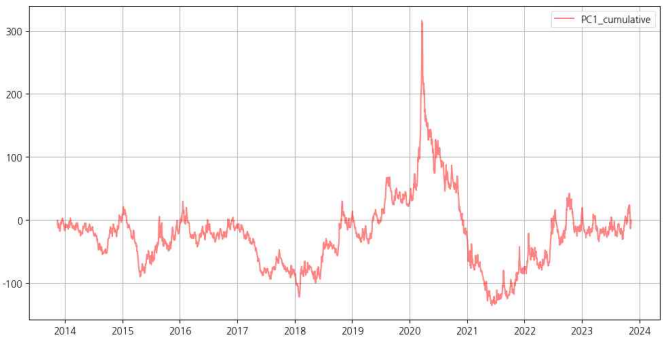
4.1.1 KOSPI 200 : 첫 번째 주성분

주성분 분석에서 첫 번째 주성분은 고윳값이 가장 크므로 전체 분산을 잘 설명할 수 있다. 수익률 데이터에서 분산은 위험으로 정의될 수 있다. 수익률에 내제되어 있는 위험을 가장 잘 설명할 수 있는 첫 번째 주성분은 그림 3과 같다.

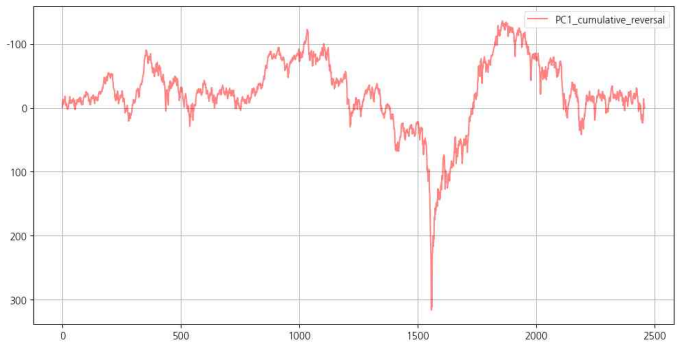


<그림3 : KOSPI200, 첫 번째 주성분의 시계열>

추세를 파악하기 위한 PC1의 누적 시계열은 그림4와 같고, 이를 반전시키면 그림5와 같다.



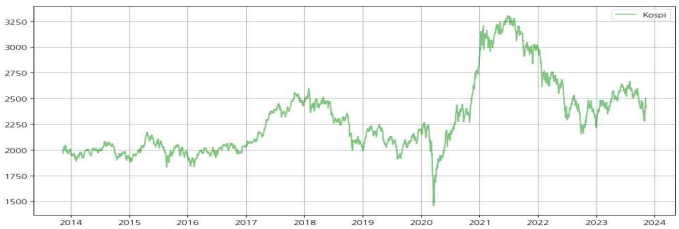
<그림4 : KOSPI200, 첫 번째 주성분의 누적 시계열>



<그림5 : KOSPI200, 첫 번째 주성분의 누적 시계열 반전>

보통 요인 모형의 경제학적 접근에서 첫 번째 요인은 시장 영향이라고 여겨진다. 위 분석에서 찾아낸 첫 번째 주성분 즉 첫 번째 요인을 시장 영향이라고 볼 수 있는 근거를 다음과 같이 찾을 수 있다.

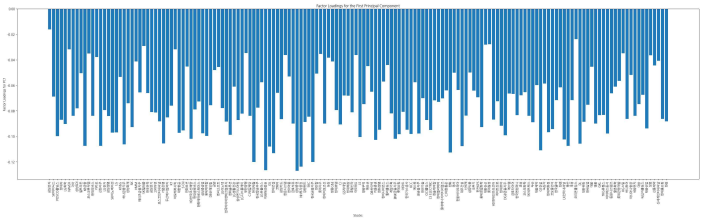
4.1.1.1 KOSPI 지수와의 유사성



<그림6 : KOSPI200, 최근 10년간 코스피지수>

그림6은 동 기간 동안의 KOSPI 지수이다. 그림6과 첫 번째 주성분의 누적 시계열을 반전시킨 그림5가 비슷한 형태가 보임을 확인할 수 있다.

4.1.1.2 베타



<그림7 : KOSPI200, 첫 번째 주성분에 대한 베타>

요인 모형으로 추정한 첫 번째 주성분에 대한 베타는 그림7과 같다. 첫 번째 주성분에 대해서 모든 주식이 같은 방향으로 반응하고 있음을 확인할 수 있다.

4.1.1.3 국내 지수와의 상관관계

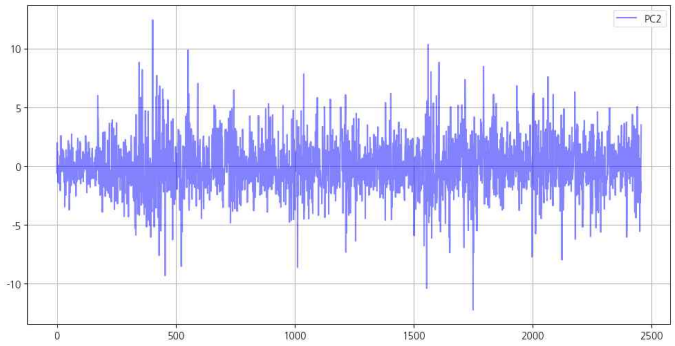
지수	상관계수
코스피 중형주	-0.79
복합 산업	-0.78
기계	-0.74
경기소비재	-0.74
증권	-0.74

<표1 : KOSPI200, 첫 번째 주성분의 누적시계열과 국내 지수의 상관계수>

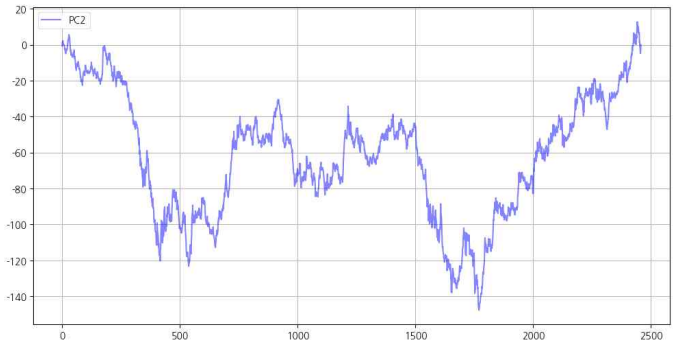
에너지, 소재, 산업재, IT 등의 총 103개의 국내 지수와 첫 번째 주성분의 누적 시계열과의 상관관계를 내림차 순한 결과는 표1과 같다. 상위 5개만 표시하였다.

4.1.2 KOSPI 200 : 두 번째 주성분

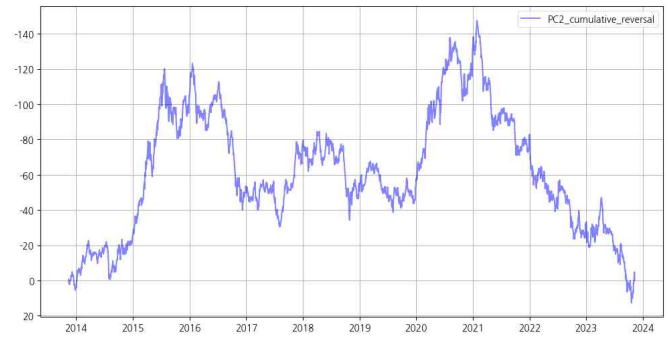
주성분 분석에서 두 번째 주성분은 고윳값이 두 번째로 크다. 두 번째 주성분의 시계열, 누적 시계열, 누적 시계열 반전은 각각 그림7,8,9와 같다.



<그림7 : KOSPI200, 두 번째 주성분의 시계열>



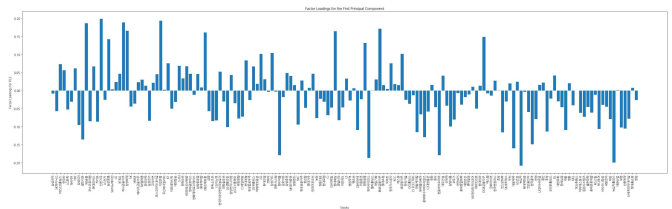
<그림8 : KOSPI200, 두 번째 주성분의 누적 시계열>



<그림9 : KOSPI200, 두 번째 주성분의 누적 시계열 반전>

두 번째 주성분 즉 두 번째 요인을 바이오섹터 라고 볼 수 있는 근거를 다음과 같이 찾을 수 있다.

4.1.2.1 베타



<그림10 : KOSPI200, 두 번째 주성분에 대한 베타>

그림10을 통해 두 번째 주성분에 대한 베타를 확인할 수 있다. 두 번째 주성분에 민감하게 반응하는 기업은 다음과 같다. [셀트리온, 신한지주, 기업은행, KB금융, 한미약품, 한미사이언스, 녹십자, 녹십자 홀딩스, DGB 금융지주, 한올 바이오파마 등] 이들 기업은 주로 바이오섹터 및 금융섹터이다.

4.1.2.2 국내지수와의 상관관계

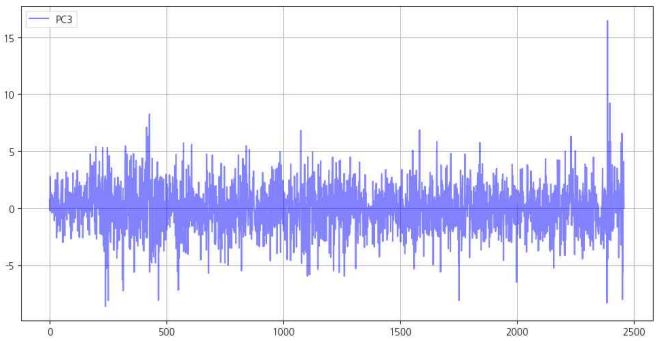
지수	상관계수
제약	-0.85
의료	-0.74
제약 및 바이오	-0.73
일반 소프트웨어	-0.66
의료 장비 및 서비스	-0.64

<표2 : KOSPI200, 두 번째 주성분의 누적시계열과 국내 지수의 상관계수>

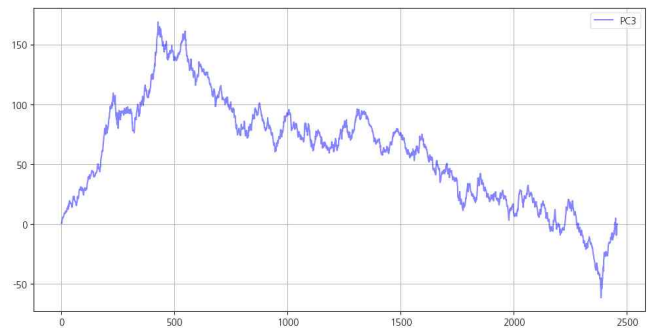
국내 지수와 두 번째 주성분의 누적 시계열과의 상관관계를 내림차 순한 결과는 표2와 같다. 상위 5개만 표시하였다. 두 번째 주성분은 바이오 관련 지수들과 상관계수가 높은 것을 확인할 수 있다.

4.1.3 KOSPI 200 : 세 번째 주성분

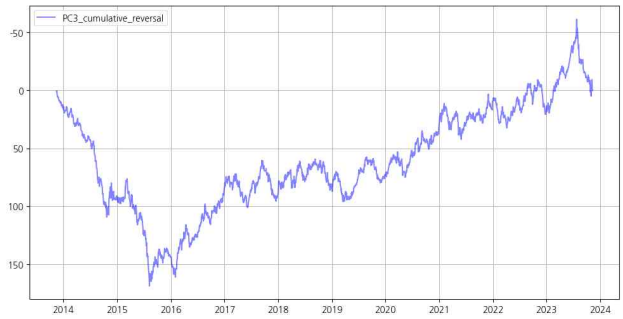
주성분 분석에서 세 번째 주성분은 고윳값이 세 번째로 크다. 세 번째 주성분의 시계열, 누적 시계열, 누적 시계열 반전은 각각 그림11,12,13과 같다.



<그림11 : KOSPI200, 세 번째 주성분의 시계열>



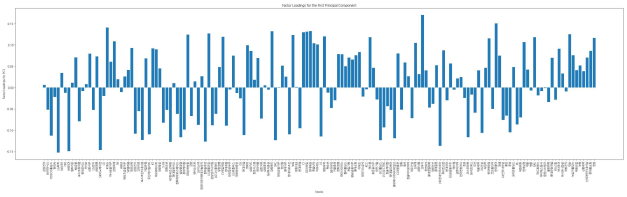
<그림12 : KOSPI200, 세 번째 주성분의 누적 시계열>



<그림13 : KOSPI200, 세 번째 주성분의 누적 시계열 반전>

두 번째 주성분 즉 세 번째 요인을 물가 및 소비라고 볼 수 있는 근거를 다음과 같이 찾을 수 있다.

4.1.3.1 베타



<그림14 : KOSPI200, 세 번째 주성분에 대한 베타>

그림14를 통해 세 번째 주성분에 대한 베타를 확인할 수 있다. 세 번째 주성분에 민감하게 반응하는 기업은 다음과 같다. [오뚜기, 하나 금융지주, HD 현대 인프라 코어, SK이노베이션, KT&G, LG화학 등] 두 번째 주성분과 달리 특정 섹터에 몰려있지 않고 골고루 분포하는 모습을 볼 수 있다.

4.1.3.2 국내지수와의 상관관계

지수	상관계수
음식료 및 담배	0.85
개인생활용품	0.85
필수소비재	0.84
전자 장비 및 기기	-0.83
식료품	0.81

<표3 : KOSPI200, 세 번째 주성분의 누적시계열과 국내 지수의 상관계수>

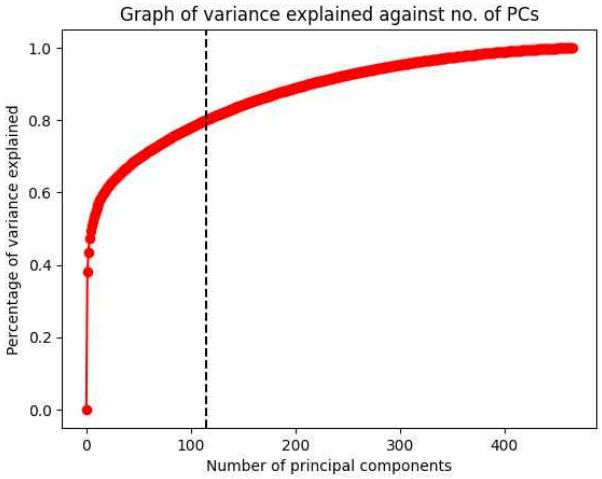
국내 지수와 세 번째 주성분의 누적 시계열과의 상관관계를 내림차 순한 결과는 표3과 같다. 상위 5개만 표시하였다. 세 번째 주성분은 물가 및 소비 지수들과 상관계수가 높은 것을 확인할 수 있다.

4.1.4 추가 분석

앞서 각각의 주성분을 추정한 방식으로 6번째 주성분 까지를 추가로 추정하면 다음과 같다. 네 번째 주성분은 IT, 다섯 번째 주성분은 금리 및 물가 그리고 여섯 번째 주성분은 상업 서비스에 해당한다고 분석했다.

4.2 S&P 500

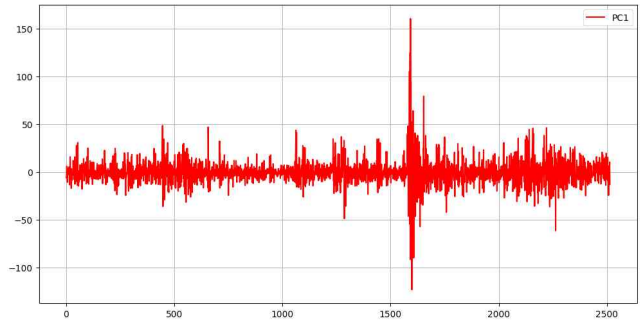
2013년 11월 13일부터 2023년 11월 10일까지의 현재 기준 S&P 500종목들의 종가를 대상으로 분석을 진행하였다. 해당 기간 내의 신규 상장 종목들은 제거를 하고 분석을 진행하였다. 종가 데이터를 수익률 데이터로 변환 후, PCA를 위해 정규화를 진행한다. 그 후, 전체 분산의 약 80%를 설명하는 115개의 주성분을 선정한다. 나머지 주성분과 고윳값의 차이가 큰 앞의 3개의 주성분만 자세히 분석하기로 한다.



<그림15 : S&P500, 고윳값을 활용한 주성분 개수 추정>

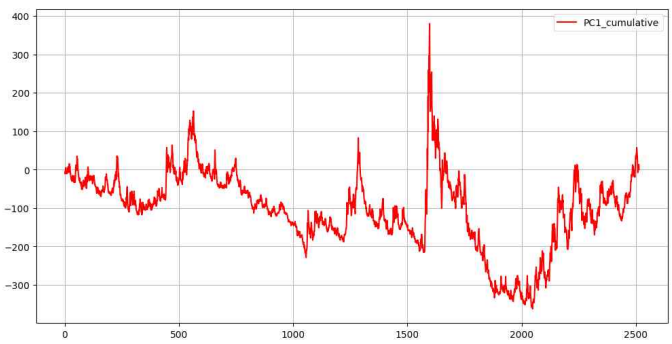
4.2.1 S&P 500 : 첫 번째 주성분

수익률에 내제되어 있는 위험을 가장 잘 설명할 수 있는 첫 번째 주성분의 시계열은 그림17과 같다.

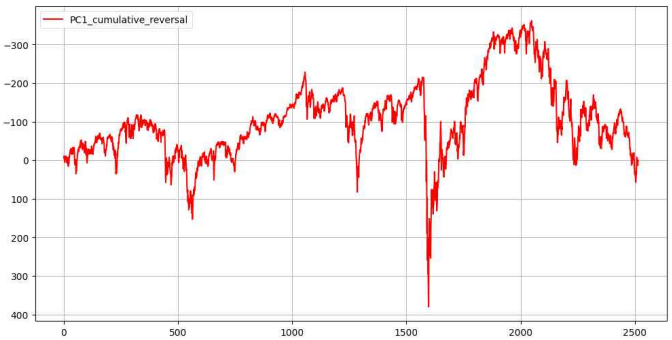


<그림16 : S&P500, 첫 번째 주성분의 시계열>

추세를 파악하기 위한 PC1의 누적 시계열은 그림17과 같고, 이를 반전시키면 그림 18과 같다.



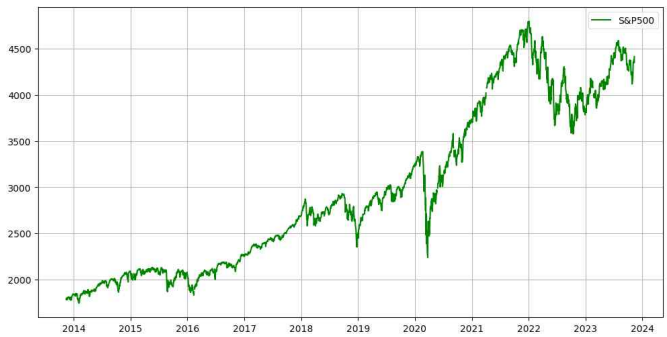
<그림17 : S&P500, 첫 번째 주성분의 누적 시계열>



<그림18 : S&P500, 첫 번째 주성분의 누적 시계열 반전>

주성분 즉 첫 번째 요인을 시장 영향이라고 볼 수 있는 근거를 다음과 같이 찾을 수 있다.

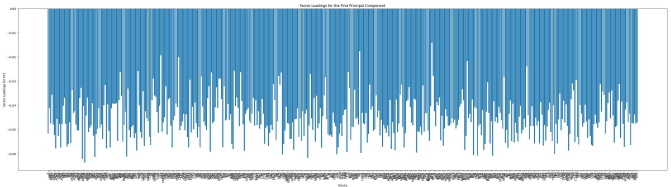
4.2.1.1 S&P500 지수와의 유사성



<그림19 : S&P500, 최근 10년간 S&P500지수>

4.2.1.2 S&P500 : 베타

그림19은 동 기간 동안의 S&P500이다. 그림19 첫 번째 주성분의 누적 시계열을 반전시킨 그림18이 비슷한 형태가 보임을 확인할 수 있다.



<그림20 : S&P500, 첫 번째 주성분에 대한 베타>

요인 모형으로 추정한 첫 번째 주성분에 대한 베타는 그림20과 같다. 첫 번째 주성분에 대해서 모든 주식이 같은 방향으로 반응하고 있음을 확인할 수 있다.

4.2.1.3 S&P500 : 미국지수와의 상관관계

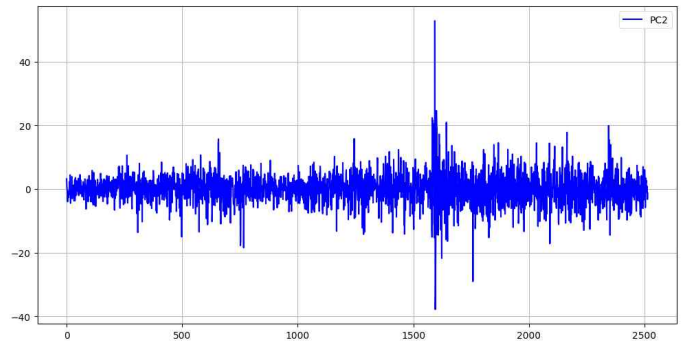
지수	상관계수
다우 소형주	-0.93
다우 윌셔 미국 소형주	-0.93
다우 윌셔 미국 중형주	-0.93
다우 전자기기	-0.91
다우 특수광물	0.91

<표4 : S&P500, 첫 번째 주성분의 누적시계열과 미국 지수의 상관계수>

다우 철강, 다우 종합TR, 나스닥 종합등의 총 172개의 미국 지수와 첫 번째 주성분의 누적 시계열과의 상관관계를 내림차 순한 결과는 표4과 같다. 상위 5개만 표시하였다.

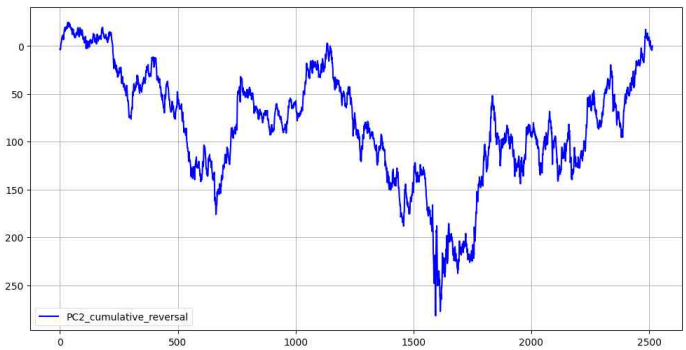
4.2.2 S&P500 : 두 번째 주성분

두 번째 주성분의 시계열, 누적 시계열, 누적 시계열 반전은 각각 그림21,22,23과 같다.



<그림21 : S&P500, 두 번째 주성분의 시계열>

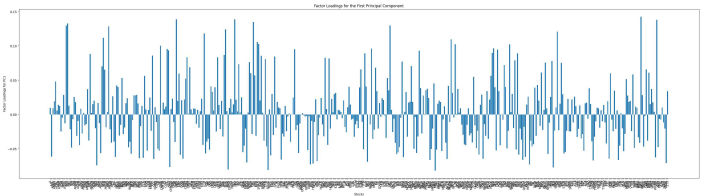
<그림22 : S&P500, 두 번째 주성분의 누적 시계열>



<그림23 : S&P500, 두 번째 주성분의 누적 시계열 반전>

두 번째 주성분 즉 두 번째 요인을 유가라고 볼 수 있는 근거를 다음과 같이 찾을 수 있다.

4.2.2.1 S&P500 : 베타



<그림24 : S&P500, 두 번째 주성분에 대한 베타>

그림10을 통해 두 번째 주성분에 대한 베타를 확인할 수 있다. 두 번째 주성분에 민감하게 반응하는 기업은 다음과 같다.

[WEC Energy Group Inc, Consolidated Edison Inc, CMS Energy Corporation, Xcel Energy Inc, Eversource Energy 등] 이들 모두는 에너지 관련 기업입니다.

4.2.2.2 S&P500 : 미국지수와의 상관관계

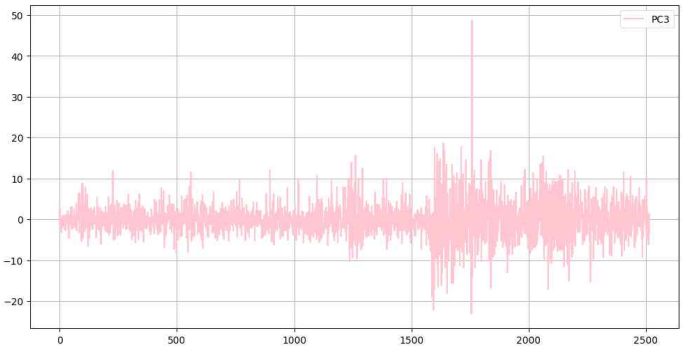
지수	상관계수
다우 오일장비&공급	-0.80
다우 에너지	-0.80
다우 오일장비&서비스	-0.79
다우 운송관	-0.72
다우 탐사&제조	-0.72

<표5 : S&P500, 두 번째 주성분의 누적시계열과 미국 지수의 상관계수>

국내 지수와 두 번째 주성분의 누적 시계열과의 상관관계를 내림차 순한 결과는 표5와 같다. 상위 5개만 표시하였다. 두 번째 주성분은 유가 관련 지수들과 상관계수가 높은 것을 확인할 수 있다.

4.2.3 S&P500 : 세 번째 주성분

세 번째 주성분의 시계열, 누적 시계열, 누적 시계열 반전은 각각 그림25,26,27과 같다.



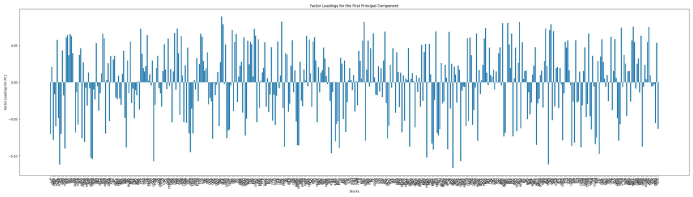
<그림25 : S&P500, 세 번째 주성분의 시계열>

<그림26 : S&P500, 세 번째 주성분의 누적 시계열>

<그림27 : S&P500, 세 번째 주성분의 누적 시계열 반전>

두 번째 주성분 즉 세 번째 요인을 에너지라고 볼 수 있는 근거를 다음과 같이 찾을 수 있다.

4.2.3.1 S&P500 : 베타



<그림28 : S&P500, 세 번째 주성분에 대한 베타>

그림28을 통해 세 번째 주성분에 대한 베타를 확인할 수 있다. 세 번째 주성분에 민감하게 반응하는 기업은 다음과 같다. [ServiceNow Inc, Synopsys Inc, Adobe Inc, Cadence Design Systems Inc, NVIDEA Corporation 등] 이들은 IT 관련 기업들입니다. 두 번째 주성분과는 달리 반응하는 정도가 기업별로 상대적으로 비슷하다고 볼 수 있다.

4.2.3.2 S&P500 : 미국지수와의 상관관계

지수	상관계수
다우 에너지 가공	0.74
다우 오일가스 생산	0.65
다우 종합보험	0.65
다우 에너지	0.59
다우 탐사&제조	0.55

<표6 : S&P500, 세 번째 주성분의 누적시계열과 미국 지수의 상관계수>

미국 지수와 세 번째 주성분의 누적 시계열과의 상관관계를 내림차 순한 결과는 표6과 같다. 상위 5개만 표시하였다. 세 번째 주성분은 에너지 관련 지수들과 상관계수가 높은 것을 확인할 수 있다.

4.2.4 추가 분석

앞서 각각의 주성분을 추정한 방식으로 6번째 주성분 까지를 추가로 추정하면 다음과 같다. 네 번째 주성분은 코로나19, 다섯 번째 주성분은 에너지 + 미디어, 방송 그리고 여섯 번째 주성분은 반도체에 해당한다고 분석했다.

4.3 KOSPI200과 S&P 500간의 비교

KOSPI200 PC	추정요인
K-PC1	시장영향
K-PC2	바이오
K-PC3	물가 + 소비
K-PC4	IT
K-PC5	금리 + 물가
K-PC6	상업서비스
S&P500 PC	추정요인
S-PC1	시장영향(특히, 중 소형주의 영향이 크다)
S-PC2	유가
S-PC3	에너지
S-PC4	코로나
S-PC5	에너지 + 미디어, 방송
S-PC6	반도체

한국 시장에서는 상위 주성분이 주도주와 거시경제변수들로 이루어져 있는 것을 확인할 수 있다. 하지만 미국은 이와는 다르게 거시경제 변수와 세계적인 사건 위주로 상위 주성분이 분포하고 있는 것을 확인할 수 있다.

5. 결론

본 고에서는 지난 10년(2013.11.13. ~ 2023.11.12.)간 KOSPI200과 S&P500 수익률 데이터에 주성분 분석을 적용해 상위 주성분을 추출한 후, 주성분의 의미를 파악하였다. 주성분의 의미를 파악하기 위해서 각 주성분에 대한 주식들의 베타, 지수와의 상관관계를 사용하였다. 그 후 KOSPI200과 S&P500의 주요 주성분을 비교해 각각의 특징을 도출하였다.

5. 참고 자료

[1] Facotor Models, MIT OpenCourseWare, Topics in Mathmatics with Applications in Finance
[2] Factor Models, Machine Learning, and Asset Pricing, Annual Review of Financial Economics, Stefano Giglio, Bryan Kelly, Dacheng Xiu, 2022

6. 알림