글로벌퀀트 R을 활용한 퀀트전략-(2) 국면 유사도 분석

Global Quant
박우열
☎ (02) 3772-2324
✉ wooyeol.park@shinhan.com

- 국면별 스타일 성과차 설명에 유용한 국면 유사도 분석
- 다양한 유형의 국면 유사도 전략: 경기 / 계절성 / 물가
- 머신러닝을 통한 국면 분류 방법: 경제지표 클러스터링

국면 유사도 분석이 유용한 이유

국면별 스타일 성과가 다르기 때문에 현재 국면과 유사했던 과거 경험치 유용

당사는 22년도 전망에서 경기순환국면에 근거해 앞으로 다가올 국면을 제시했다. 각 국면별로 강세를 보이는 스타일과 자산군이 다르기에 유사했던 과거 경험을 근거로 자산을 선별할 때 매우 유용하기 때문이다. 경기 뿐 아니라 이익, 물가, 금리 등 다양한 지표를 토대로 국면을 분할하려는 시도가 계속되는 이유가 이것 때문이다.

다양한 국면구분 방법

본 보고서에서는 ① 경기순환국면 ② 계절성국면 ③ 물가국면에 대한 분석을 토대로 현재와 유사했던 과거 국면에서 유망했던 스타일과 자산군이 무엇인지를 밝혔다. 최종적으로는 국면 분할에 사용하는 수 많은 변수들을 모두 종합해 머신러닝을 적용했다. 즉, 사람이 한 번에 판단하기에는 변수가 너무 많다면 기계가 대신 현재 국면과 유사했던 과거 국면을 종합적으로 판단하도록 해봤다.

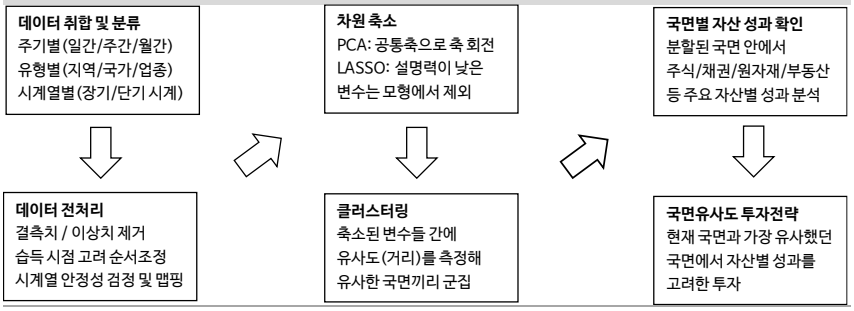
분석에 사용한 통계 기법

R을 활용하면 머신러닝 뿐 아니라 다양한 통계 기법의 사용이 수월해지는데 본 보고서에서는 주성분분석(PCA), LASSO회귀, 계절성 분석, 클러스터링 기법을 활용했고 각 기법의 원리와 구현을 위한 R 코드는 부록으로 정리했다.

투입 변수의 개수가 많아질수록 데이터의 입수와 결측치, 이상치, 시계열 안정성에 대한 전처리 및 유효한 변수를 선별하는 일이 중요해진다. 머신러닝 학습에 변수를 투입할 때 필요한 각종 데이터 전처리 기법도 부록으로 정리했다.

국면 유사도 투자전략

국면유사도 활용 투자전략 도식



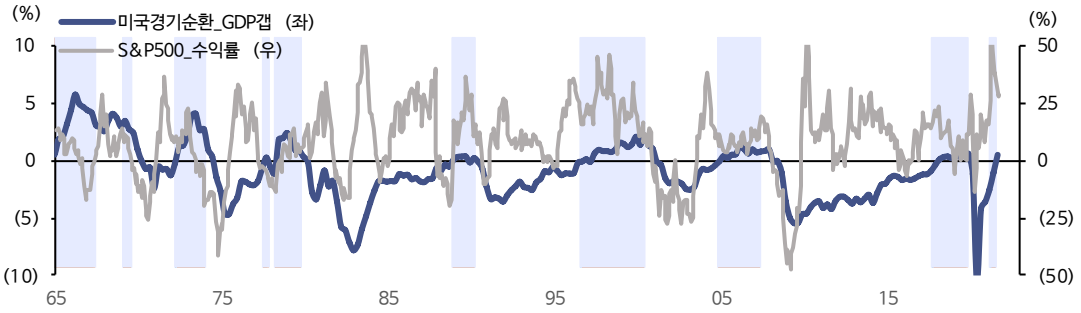
자료: 신한금융투자

다양한 유형의 국면유사도 투자전략: ① 경기순환주기 국면

국면을 구분하는 한 가지 방법은 GDP갭의 사용

경기순환주기로 국면을 분할하는 한 가지 방법은 실제 GDP와 잠재GDP의 차이로 계산한 GDP갭을 사용하는 것이다. GDP갭이 (+)인 시기에는 실업률이 낮고 총수요가 총공급보다 많아 경기가 확장국면에 있는 상황이다. GDP갭의 레벨값과 기울기를 사용해 4개 국면으로 구분한 뒤 각 국면을 초반과 후반으로 추가로 분할해 총 8개로 국면을 분할하였다.

경기순환주기를 활용한 국면 구분 및 현재와 유사했던 국면에서 증시 성과



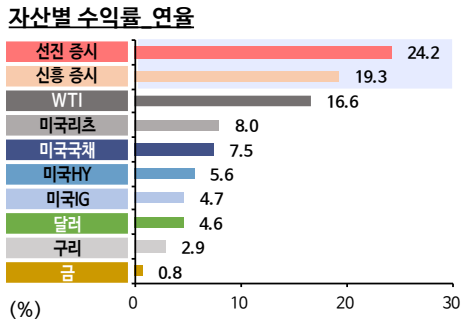
자료: Refinitiv, Bloomberg, 신한금융투자

경기순환주기 상 현재는 Mid-Late Cycle 구간이고 이 때 수익률이 좋았던 자산은 선진국 주식

유휴자원 없이 모든 설비시설이 작동하고 실업률이 매우 낮아 경기가 확장하는 국면에서는 민감주가 방어주보다 낮고 리스크 수용도가 높아 채권보다 주식이 낫다. 실제로 과거 경험이 그러했는지 확인하기 위해 지금부터 내년도까지 우리가 지나갈 경로와 동일했던 국면에서 각 자산별 성과를 실증분석했다.

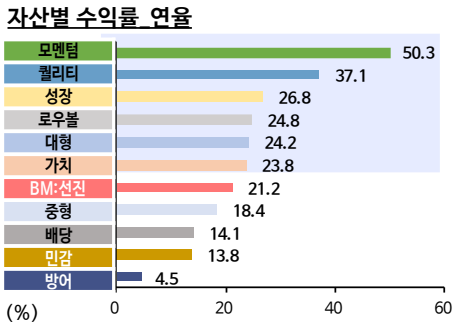
국면 분류와 성과는 1980년 이후 데이터를 사용했다. 그 결과 다양한 자산군 중에서도 주식의 성과가 좋았고 그 중에서도 선진국, 선진국 주식 중에서도 모멘텀과 퀄리티 성과가 우수했다.

경기순환주기 Mid-Late Cycle 자산별 수익률



자료: Refinitiv, Bloomberg, 신한금융투자

경기순환주기 Mid-Late Cycle 스타일별 수익률



자료: Refinitiv, Bloomberg, 신한금융투자

경기순환주기 상 유사 국면 과거 자산별 성과

확장 1국면				확장 2국면				둔화 1국면		
01/89-04/89	12/96-05/98	02/05-08/05	01/18-11/18	05/89-09/89	06/98-11/99	09/05-02/06	12/18-11/19	09/89-12/89	12/99-07/00	03/06-02/07
WTI (35.6)	선진증시 (39.2)	WTI (30.0)	WTI (38.7)	신흥증시 (16.5)	WTI (56.5)	구리 (33.7)	금 (24.8)	신흥증시 (24.5)	WTI (23.0)	구리 (38.9)
신흥증시 (8.7)	미국리츠 (22.0)	신흥증시 (11.9)	선진증시 (9.7)	WTI (10.7)	선진증시 (16.9)	신흥증시 (26.0)	미국리츠 (20.6)	구리 (12.9)	신흥증시 (11.5)	신흥증시 (32.4)
선진증시 (7.6)	달러 (15.7)	미국리츠 (11.0)	미국HY (1.4)	미국국채 (10.5)	미국국채 (6.1)	금 (19.7)	미국IG (18.0)	선진증시 (8.0)	달러 (11.3)	미국리츠 (30.9)
달러 (2.2)	미국국채 (12.5)	달러 (10.5)	달러 (0.5)	미국리츠 (6.9)	구리 (2.3)	선진증시 (10.7)	미국국채 (11.1)	미국국채 (2.3)	선진증시 (10.4)	금 (23.8)
미국국채 (1.5)	신흥증시 (-6.7)	구리 (4.2)	미국국채 (-1.8)	달러 (5.5)	금 (1.9)	WTI (3.3)	미국HY (9.0)	금 (-0.4)	미국리츠 (8.5)	선진증시 (20.5)
미국리츠 (1.3)	구리 (-9.7)	선진증시 (2.2)	미국IG (-3.8)	선진증시 (-1.9)	달러 (-3.7)	미국리츠 (2.9)	달러 (4.0)	WTI (-1.0)	미국HY (5.2)	미국HY (7.9)
구리 (-3.2)	금 (-20.5)	미국국채 (1.8)	미국리츠 (-4.6)	금 (-2.3)	신흥증시 (-6.7)	미국HY (1.4)	선진증시 (1.2)	달러 (-2.4)	미국국채 (4.5)	미국IG (4.1)
금 (-6.6)	WTI (-35.6)	미국HY (1.4)	신흥증시 (-4.8)	구리 (-22.4)	미국리츠 (-27.5)	달러 (1.2)	신흥증시 (-1.7)	미국리츠 (-5.3)	미국IG (1.8)	미국국채 (3.2)
		미국IG (0.9)	금 (-6.9)			미국국채 (0.7)	구리 (-9.6)		구리 (0.6)	WTI (0.0)
		금 (-5.4)	구리 (-9.3)			미국IG (-0.3)	WTI (-28.8)		금 (-11.1)	달러 (-8.2)

자료: Refinitiv, Bloomberg, 신한금융투자

그런데 1) GDP 갭을 단일 독립변수로 사용한 것과 2) 국면 구분방법으로 변수의 부호(+/-)와 기울기를 사용한 것은 경기순환주기라는 경제학적 모델을 근거로 한 연역적 방법이다.

이것은 물론 타당한 방법이지만 수많은 거시경제지표 중 어떤 것을 사용해야 "스타일별 성과 차이를 가장 잘 구분해줄 수 있는 국면"을 찾을 수 있을지 고민한다면 다른 접근이 필요하다. 이 관점에서 보면 GDP갭을 사용한 국면분할은 무수히 많은 분할 방법 중의 한 가지 특수한 형태였을 뿐이다.

다양한 유형의 국면유사도 투자전략: ② 계절적 국면

계절적 반복성에 근거한 투자도 국면유사도 전략

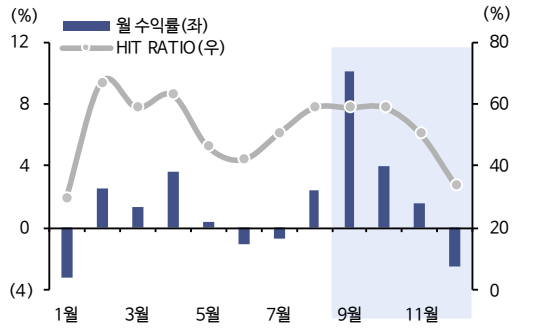
계절적으로 특정 산업의 반복적인 변동을 포착하는 것도 국면 유사도 투자전략의 일종이다. 겨울에 추워지거나 연말 쇼핑 시즌에 소비가 늘어나는 것은 매년 발생하는 예측 가능한 패턴이기 때문이다. 계절성으로 널리 알려진 사실들이 있는데 다음과 같은 몇 가지이다.

계절성 분석 예시

(주장)겨울에는 난방수요가 올라가서 천연가스 수요가 늘어날 것이기에 가격이 오를 것이다. (근거) 97년1월-20년12월 데이터로 천연가스 월 평균 수익률을 확인했더니 실제로 9월-11월 사이 천연가스 가격이 올라가는 계절성이 있었다.

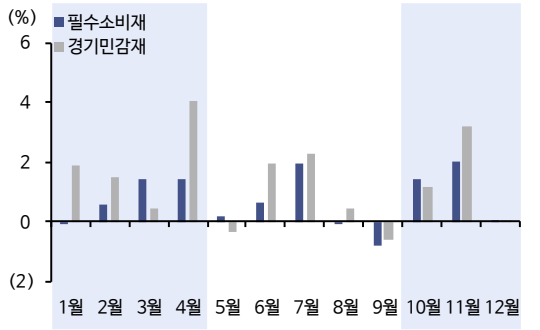
(주장) 연말 쇼핑시즌에 소비가 늘어나면 리테일 업종 수익성이 개선되기 때문에 성과가 더 나을 것이다. (근거) 미국 소매판매 업종의 월 평균 수익률을 확인했더니 10월부터 다음해 4월까지 성과가 양호했다.

천연가스 가격 계절성



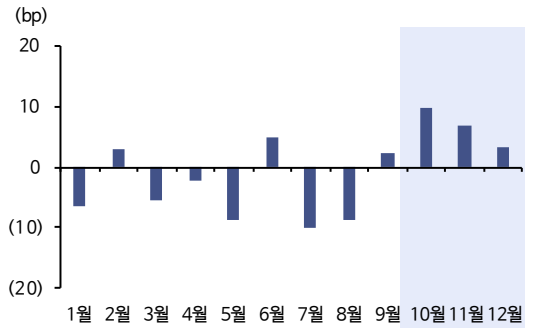
자료: Fred Economic Research, R, 신한금융투자
주) 1997년 1월 - 2020년 12월 데이터

미국 소매판매 업종 월별 수익률 계절성



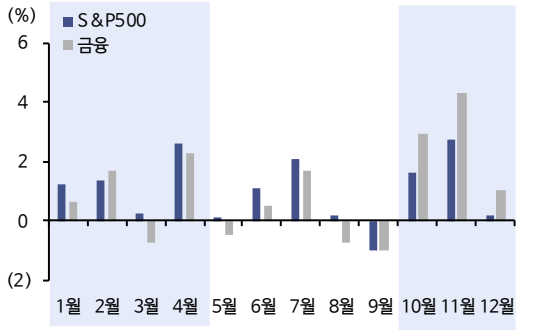
자료: Refinitiv, R, 신한금융투자
주) GICS 11개 업종 분류, 2011년 1월 이후 데이터

미국 10년물 금리 계절성



자료: Refinitiv, R, 신한금융투자
주) 2011년 1월 - 2021년 10월 데이터

미국 업종별 성과 계절성



자료: Refinitiv, R, 신한금융투자
주)GICS 분류 11개 업종, 2011년 1월 - 2021년 10월 데이터

계절적 증시 변동을 설명하기 위한 노력

S&P 500의 월 성과 계절성을 보면 "9월에 사서 5월에 팔아라"는 월가 격언이 확인이 되고 있다. 데이터로 확인되더라도 계절적 요인이 어떻게 주가로 이어지는지 그 연결고리가 불분명하기에 이를 설명하기 위한 노력도 다양하다.

2003년 애틀란타 연방은행은 "Winter Blues: A SAD Stock Market Cycle"이라는 조사분석자료에서 계절성 정서장애(SAD:Seasonal Affective Disorder)라는 임상적으로 널리 알려진 사실을 이용해 주가계절성을 설명했다.

계절성 정서장애로 리스크 수용도 및 주식 프리미엄 또한 계절적으로 변동함

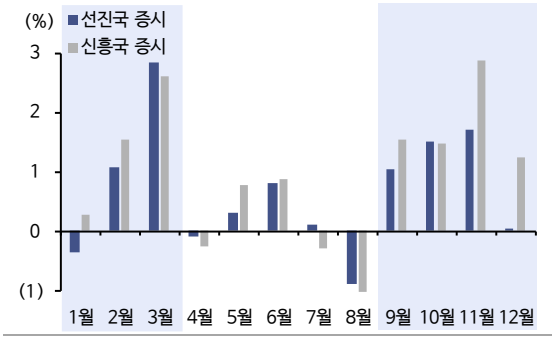
계절적 요인으로 우울감을 느끼는 사람이 많아지면 리스크 수용도가 변화해서 주가 프리미엄도 바뀔 수 있다. 적도와의 거리에 따른 나라별 기후분포 차이로 계절적으로 따른 일조량 변동이 큰 나라, 즉 위도값이 큰 북반구 지역의 주식시장일수록 계절적 요인으로 인한 주식시장 수익률 변동이 크다는 것이다.

위 연구결과는 2003년 기준이기에 지금도 나라별 증시 계절성이 존재하는지 추가 연구를 진행했다. 1999년 1월부터 2021년 10월까지의 증시 성과를 지역별로 구분했다. 11월부터 투자를 고려한다면 선진국보다 신흥국 성과가 더 나았다. 나라별 증시성과의 계절적 차이를 설명할 수 있는 또다른 설명은 다음과 같다.

계절적 수혜를 받는 특정 산업이 있을 수 있고 이 산업이 중요한 국가가 있을 수 있음

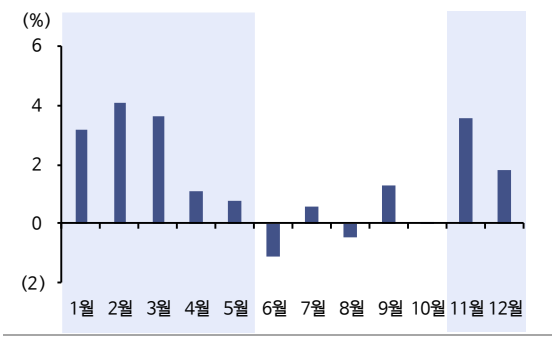
천연가스는 겨울철 난방수요 증가로 가격 상방압력을 받으므로 이를 수출하는 나라의 증시 이익 추정치에도 계절성이 있을 것이다. 2020년말 기준 천연가스 최대 수출국은 러시아로 약 2400억 입방미터를 수출했는데 러시아 증시의 달러 환산 성과도 겨울철에 들어가기 직전인 11월부터 성과 개선이 뚜렷했다.

글로벌 증시 월간 수익률 계절성



자료: Refinitiv, R, 신한금융투자
주) MSCI 지수, 1999년 1월 - 2021년 10월 데이터

러시아 증시 월간 수익률 계절성(달러환산)



자료: Refinitiv, R, 신한금융투자
주) MSCI 러시아 지수, 1999년 1월 - 2021년 10월 데이터

다양한 유형의 국면유사도 투자전략: ③ 물가 국면

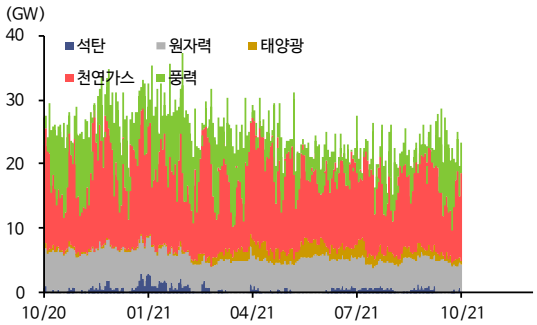
물가 국면 분석

인플레이션이 일시적인 것인지 아니면 계속될 것인지에 대한 논쟁이 뜨겁다. 친환경 규제와 맞물린 원자재 공급망 변화로 각종 원자재 물가가 올라가고 있다. 소위 그린플레이션으로 불리는 현상이며 이는 영국 에너지원별 발전량을 보면 확인된다. 올해 9월달 유럽은 바람이 많이 안불었고 풍력이 안정적인 에너지 공급원으로 역할을 못하자 계절적으로 가장 비싸지는 천연가스 의존도를 높일 수 밖에 없었다.

물가 상승이 증시에 영향을 미치는 경로는 다양하다. 원가상승을 판가에 얼마나 잘 전가시킬 수 있는지에 따라 산업별 영향이 다르다. 생산자물가(PPI) 상승분을 소비자물가(CPI) 상승분으로 전가시키는 능력에 따라 산업별 수혜가 달라진다. 또한, 물가는 중앙은행의 금리인상과 테이퍼링의 시점이 언제가 될지에 대한 중요한 변수로 작용한다. 금리가 올라가는 시기는 밸류에이션 멀티플 확장이 어려운 국면이다.

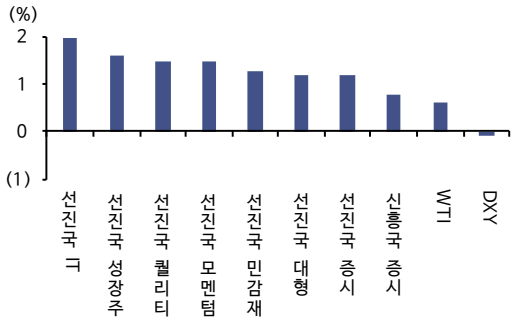
미국 PCE 단일지표를 사용해 PCE 레벨 값과 기울기 값을 기준으로 4개 국면으로 구분해보았다. 이번 물가 국면에서 자산별 성과는 다음과 같았다. 주로 선진국 증시가 강세를 보였으며 그 중에서도 IT, 성장주, 퀄리티는 월 평균 2% 수준의 성과를 보였다.

영국 발전원별 발전량



자료: UK National Grid, 신한금융투자

이번 물가 국면에서 과거 자산별 성과



자료: Refinitiv, R, 신한금융투자

다양한 유형의 국면유사도 투자전략: ④ 머신 러닝을 사용한 통합

머신러닝을 이용한 국면유사도 측정

앞에서 살펴본 것 이외에도 국면을 분할할 때 유용한 것으로 널리 알려진 변수가 매우 많다. 이렇게 일간, 주간, 월간으로 발표되는 변수의 개수는 이제 사람의 직관으로 한 번에 파악하기 힘들다. 이 변동을 모두 사용해서 지금 국면과 가장 유사한 과거 국면을 찾을 수 있다면 어떨까? 사람의 직관만을 사용하면 매우 힘든 일이지만 머신러닝 기법을 적용하면 이것이 가능하다.

군집분석에 사용하는 머신러닝 기법 중 k-means 클러스터링 기법을 사용해 국면을 세부적으로 재분할했다. 알고리즘이 직관적이고 구현이 어렵지 않으며 국면간 계층이 없어 국면별 자산 성과를 비교하기 용이하기 때문이다. 여기서 k는 시계열을 몇 개의 군집으로 분할하는지 결정해주는 값인데 오차항을 줄여주는 최적 군집의 개수와 통상 매크로 국면을 4개 정도로 분할하는 관행을 고려하여 k=4를 적용했다. R에서 k-means 클러스터링이 군집을 4개로 분할해주는 알고리즘은 다음과 같다.

k-means 클러스터링 알고리즘

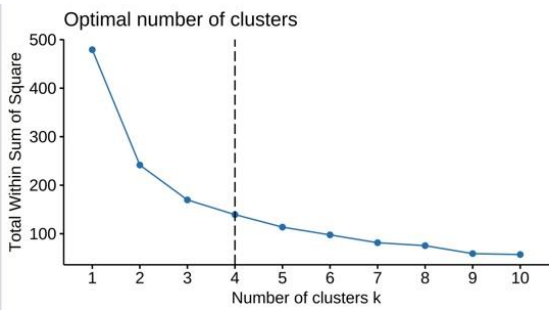
1) 임의로 점 4개(x1,x2,x3,x4)를 선택하여 각각 1~4개의 군집으로 분류한다. 각 군집에 추가적으로 데이터가 할당된다면 E(xi)는 i째 군집의 평균값으로 정의한다.

2) 다음으로 선택된 데이터(xj)에 대해 d(E(xi)-xj), i=1,2,3,4를 계산한 뒤 이 네 개 값 중 거리가 가장 짧은 군집으로 xj를 분류한다. d(_)는 거리 측도 함수인데 가장 직관적인 평면 상의 직선 거리부터 절대값 거리, 상관관계 거리 혹은 시계열 특성을 반영한 DTW(Dynamic Time Warping) 기법 적용이 가능하다.

3) 군집으로 나누고자 하는 모든 데이터에 이를 반복한다. 즉, n개의 관측치에 대해 j=5,…,n까지 이를 반복하면 모든 데이터는 4개의 군집으로 분류된다.

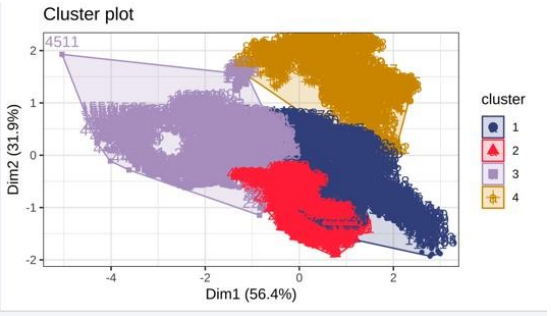
위 알고리즘은 최초에 선택하는 임의의 점이 무엇이냐에 따라 결과값이 다르다. 따라서 위 알고리즘을 1,000회 반복하고 그 중에 가장 분산이 낮은 결과를 사용했다. 길게 설명했지만 원리는 어렵지 않고 무엇보다 R에는 클러스터링 패키지가 존재하기 때문에 한 줄로 구현이 된다.

k-means 클러스터링 군집수 최적화



자료: Refinitiv, R, 신한금융투자

k-means 클러스터링을 사용한 데이터 분류



자료: Refinitiv, R, 신한금융투자

국면유사도를 활용한 투자전략

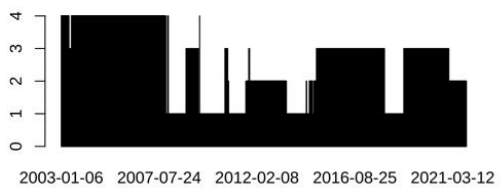
클러스터링 투입변수 선정

일간, 주간, 월간 데이터를 각각 군집으로 묶은 뒤 국면별 자산 성과를 다음과 같이 정리했다. 주간이나 일간으로 군집을 묶어도 국면의 변동이 그렇게 자주 일어나지는 않았다.

일간 지표 사용 국면 구분

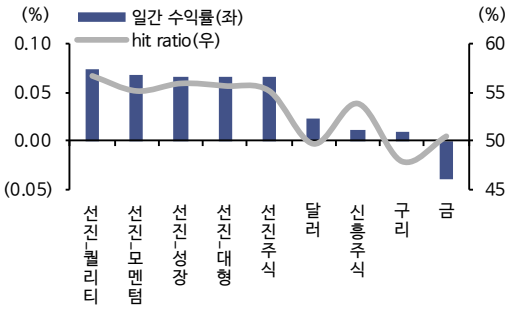
아래는 WTI, BEI 10년물, 미국 10년물 금리라는 3개 변수를 일간 데이터로 사용해 현재 국면과 같았던 구간에서 자산별 성과를 정리한 것이다. 일간 데이터를 사용해 현재 국면과 유사했던 국면에서 자산별 성과를 보면 선진국 주식이 강세를 보였고 그 중에서도 퀄리티, 모멘텀, 성장주가 강세였다.

매크로 국면 군집결과



자료: R, Fred Economic Research, 신한금융투자
주) 군집 결과로 1,2,3,4 중 하나의 숫자를 부여받는 것임

현재와 동일한 국면에서 과거 자산 성과

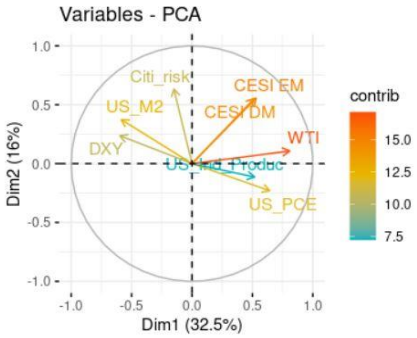


자료: R, Fred Economic Research, 신한금융투자
주)2003년 1월 ~ 2021년 10월, 4911일 영업일 중 군집된 774일

월간 지표 국면 구분

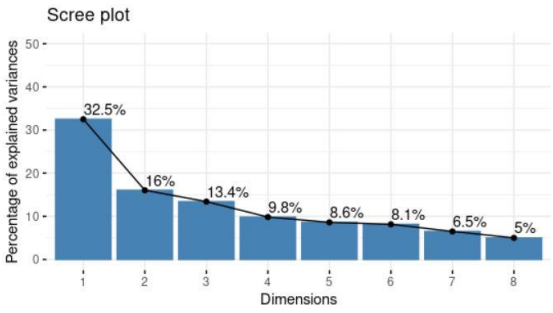
다음으로, 월간으로 발표되는 주요 매크로 지표들을 묶어 보았다. 투입 변수는 CESI, US_PCE, US_M2, US_산업생산, Citi 리스크지표, 달러, 유가를 사용했으며 클러스터링을 위해 12개월 이동평균을 사용해 표준화하였다. 투입 변수의 선정 기준과 자세한 전처리 방법 및 통계기법 원리 설명은 부록으로 별첨하였다.

PCA 차원축소 방향



자료: Refinitiv, R, 신한금융투자

PCA 공통축 Scree Plot



자료: Refinitiv, R, 신한금융투자

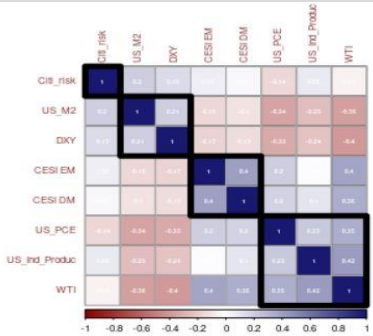
투입변수 PCA 결과

투입 변수들의 주성분분석(PCA) 분해 결과는 위와 같았다. 1번 공통축의 설명분이 32.5%, 2번 공통축의 설명분이 16%, 3번 공통축의 설명분이 13.4%였고 위 표에서 알 수 있듯 각 변수들의 변환 축 위에서의 방향은 서로 크게 겹치지 않았다.

투입변수 간 상관관계 및 클러스터링 결과

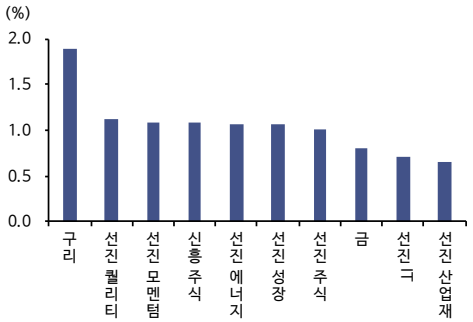
상관관계 플롯을 보면 서로 상관관계가 높은 조합은 (US M2, 달러), (CESI EM, CESI DM), (US PCE, US 산업생산, WTI)라는 세 묶음이었다. 부록에 설명한 전처리 기법을 적용해 클러스터링으로 묶은 뒤 현재 국면과 같았던 국면에서 자산별 성과를 아래와 같이 정리하였다.

매크로 변수 간 상관관계 플롯



자료: R, Refinitiv, 신한금융투자

현재와 동일한 국면에서 과거 자산 성과



자료: R, Refinitiv, 신한금융투자

결론) 현재 국면과 유사했던 국면에서 유망했던 자산

가장 짧은 주기의 국면은 계절성

경기순환국면, 물가국면, 계절성국면을 고려해 현재 국면과 유사했던 국면에서 성과가 우수했던 자산을 매수하는 전략이 여전히 유효하다. 국면 중에 주기가 가장 짧은 것은 계절성으로 매 3개월마다 변하기 때문이며 겨울철 유망한 자산에 대한 분석은 유효하다.

국면별 유망 자산

머신러닝을 적용해 일간, 월간으로 군집을 형성했을 때도 현재 국면과 유사했던 국면에서 유망했던 자산의 성격이 크게 변하지는 않았다. 기존에 알려진 국면과 머신러닝을 적용한 국면에서 공통적으로 상위권에 속했던 자산군은 선진국 주식이었으며 그 중에서도 모멘텀, 퀄리티, 성장, IT업종이 강세를 보였고 원자재 중에서는 구리가 강세를 보였다. 국면별 유망자산을 아래에 도표로 정리하였다.

국면별 유망한 자산

국면 분할 방법	현재국면	유망자산	
경기순환 국면	mid-late cycle	선진 모멘텀	선진 퀄리티
계절성 국면	겨울	미국 리테일 섹터	러시아 증시
물가 국면	물가상승	선진국 IT	선진국 성장주
일간 경제지표 국면	머신러닝	선진 퀄리티	선진 모멘텀
월간 경제지표 국면	머신러닝	구리	선진 퀄리티

자료: 신한금융투자

부록1) 빅데이터 분석 유의사항

데이터에 근거한 귀납 추론이 타당하기 위한 조건

과거 데이터에 근거한 귀납 추론을 할 때는 주의할 점이 몇 가지 있다. 첫째로 애초에 독립변수를 선택할 때부터 변수 간의 인과관계를 고려해야 한다는 점이다. 합당한 연결고리가 없는 변수들을 사용해 결과를 도출해서는 안되기 때문이다. 둘째로 과거에 반복적으로 발생했다고 해서 이번에도 되풀이될 것인지에 대한 의구심이 있다는 것이다.

빅데이터 분석에 분야 지식이 중요한 이유

머신러닝을 금융 시계열에 적용할 때 주의해야 하는 것이 바로 이 지점이다. 투자의 주체는 사람이기에 사람이 해석하기 힘든 관계를 찾는 것이 타당한지 검증이 필요하다. 변수를 투입할 때 데이터 자체의 특성을 고려하지 않으면 관계가 없는데 있는 것 처럼 보이는 가성회귀(Spurious Regression)를 도출하기도 쉽다.

통계적으로 이를 피하기 위해서 시계열 분석을 하기 전에 트랜드와 계절성, 자기상관 등 패턴을 제거하게 된다. 이를 위해서는 데이터 자체의 특성과 변수들 간의 관계를 이해하는 것이 중요하다. 통계적 기법을 금융에 적용하고자 할 때 분석대상에 대한 분야 지식(Domain Knowledge)이 중요한 이유이다.

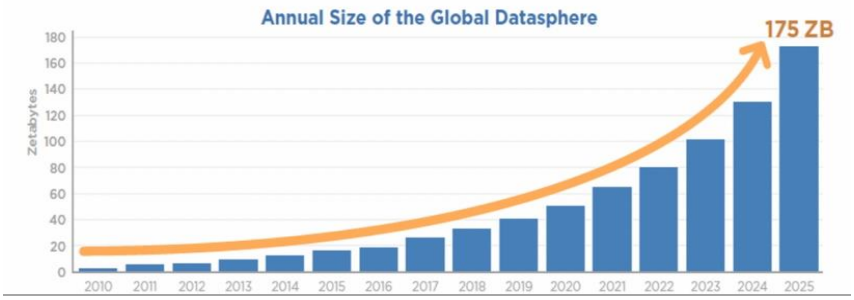
경제학적 함의가 있지만 데이터를 사용해 반복적으로 타당성을 검증할 수 없는 추론의 예시

반면에 변인 간 인과관계 제시가 합당한 추론이지만 귀납적 근거가 부족한 추론도 존재한다. 지구 온난화 때문에 사막화가 늘어나고 조업가능지역이 줄어들면 생산성이 감소할 것이고 이로 인해 아프리카처럼 더운 나라에 더 악재로 작용할 것이라고 말한다면 변수들 간의 경제학적 함의는 있다. 다만 이 경우 아직 발생하지도 않은 일인데 과거 유사한 사례에서 비슷한 결과를 반복적으로 도출할 수도 없다.

빅데이터에 근거한 투자전략도 변인 간 인과관계에 대한 설명이 필요

원유 저장고 위성사진에 대한 이미지 분석으로 원유 재고를 알아내거나 텍스트 마이닝으로 트랜디한 주식을 선별하려는 등 새로운 시도가 진행 중이다. 매일 새로 만들어지는 대체 데이터들이 급증하고 있기에 사람이 직관적으로 모두 이해하기 힘들어지고 있다. 그렇다고 기계에만 의존해 모델 결과를 무비판적으로 수용할수도 없다. 결론적으로 변인 간 인과관계가 뚜렷하고 장기간 반복되는 패턴에 근거할 수 있는 추론에 근거해야만 더 나은 투자전략이라고 할 수 있다.

급증하는 정형 / 비정형 데이터의 양



자료: IDC, 신한금융투자

부록2) 각종 데이터 전처리와 변수축소 원리 및 R 코드

국면 유사도 분석에 물가, 금리, 이익, 수급 등 특정 요소를 먼저 고려한다면 이는 어느정도 귀납적 추론에 근거하는 휴리스틱한 방법이다. 스타일별 성과에 영향력을 갖는 것으로 사전적으로 알려진 변수들을 취사선택하는 것이기 때문이다.

클러스터링에 사용할 투입변수 가능 풀

종속변수	독립변수	주기	설명	독립변수	주기	설명
R(DM-EM)	WTI	D	WTI 유가	M2	M	주요국 M2 통화량 증가율(미국,유럽,일본)
R(EQ-BND)	BEI	D	미국 BEI(Breakeven Rate)	미 PCE	M	미국 PCE(Personal Consumption Expenditures)
R(Cyc-Def)	US_10Y	D	미국 10년물 금리	미 산업생산	M	미국 산업생산 증가율
R(GRO-VAL)	DXY	D	달러인덱스	CITI_Risk	M	Citi 매크로 리스크 지표
R(LAR-SML)	CESI	D	지역별 CESI(Citi Economic Surprise Index)	PMI	M	지역별/국가별 구매관리자 서베이지수
	이익추정치	W	I/B/E/S에서 제공하는 지역별, 국가별, 섹터별 주간 이익추정치 지역은 선진/신흥/유럽/아시아, 국가는 지역별 대표국가, 섹터는 GICS 11개 업종			
	매출추정치	W	I/B/E/S에서 제공하는 지역별, 국가별, 섹터별 주간 매출추정치 지역은 선진/신흥/유럽/아시아, 국가는 지역별 대표국가, 섹터는 GICS 11개 업종			
	ROE추정치	W	I/B/E/S에서 제공하는 지역별, 국가별, 섹터별 주간 ROE추정치 지역은 선진/신흥/유럽/아시아, 국가는 지역별 대표국가, 섹터는 GICS 11개 업종			
	ETF_FF	W	지역별/업종별 ETF 수급			

자료: Refinitiv, 신한금융투자

데이터를 모은 뒤에는 결측치 및 이상치 제거 등의 전처리가 필요하며 일부 입수시점이 지연되는 변수의 경우 시점조정이 필요하다. 예를 들어 통화량이나 PMI와 같은 매크로 지표의 습득은 통상 한 달 정도 늦다.

다음으로 시계열의 안정성에 대한 검정이 필요하다. 즉, 시계열의 평균, 분산이 일정하며 공분산이 확률변수의 시차에 의존해야 한다. 시계열의 안정성을 검증하는 방법은 단위근 검정이 있는데 단위근이 있는 불안정적인 시계열이라면 차분 등을 통해 안정적으로 변환해야 한다.

R에서 제공해주는 시계열 패키지를 사용하면 몇 줄로 검증이 가능하다. Fred Economic Resaerch의 데이터를 크롤링하여 각종 전처리 및 통계 기법을 수행할 수 있도록 코드를 정리하였다. 따로 유료 데이터를 구독하지 않더라도 아래 코드를 R에서 순서대로 실행시키면 데이터를 가져와서 검정을 수행하게 된다.

천연가스 데이터 풀링 및 계절성 분해 R 코드

```
library(quantmod) # 데이터 풀링 패키지
library(xts) # 시계열 패키지(ts: time series)
library(seasonal) # 계절성 분석 패키지

natGas<-getSymbols("DHHNGSP", src="FRED", auto.assign=FALSE)
#FRED 사이트에서 천연가스 스팟가격을 크롤링
natGas_omit<-na.omit(gatGas) # 데이터 결측치 제거
natGas_omit_m<-to.monthly(natGas_omit) # 일간 데이터를 월간으로 변환
natgas_ts<-ts(natGas_omit_m$natGas_omit_Close, frequency=12, start=
c(1986,1) #데이터프레임을 시계열 데이터로 변환
natGas_adj<-seas(natGas_ts) # 시계열에 대한 계절성 분석 수행
```

자료: R, Fred Economic Research, 신한금융투자

시계열 정상성 검정(단위근 검정) R 코드

```
library(quantmod) # 데이터 풀링 패키지
library(xts) # 시계열 패키지
library(tseries) #KPSS 시계열 정상성 검정, H0: 시계열이 정상(=단위근 없음)
# ADF(Augmented Dickey Fuller) 검정, H0: 시계열이 비정상(=단위근 있음)

kpss.test(natGas_ts, "Trend") # 시계열 정상성 검정
adf.test(natGas_ts, k=1) # 시계열 정상성 검정, k는 시차의 길이(lag order)

# ADF에 사용할 k값은 잔차들이 자기상관을 가지지 않도록 선택할 것. 방법은 AIC 혹은 BIC를 최소화하는 값을 선택하거나 마지막 시차가 통계적으로 유의할때까지 시차를 바꾸는 것임
```

자료: R, Fred Economic Research, 신한금융투자

본문에서 설명한 클러스터링 알고리즘을 적용할 때 주의할 점이 또 있다. 각 변수별 평면거리를 측정하기 때문에 스케일이 서로 다른 변수를 투입하면 의도치 않게 스케일이 큰 변수에 가중치를 높게 부여하는 일이 발생한다. 따라서 투입변수를 (-1,1) 사이의 값으로 맵핑(Mapping)하여 사용하였다.

매크로 데이터 풀링 및 맵핑 R 코드

```
library(quantmod) # 데이터 풀링 패키지
library(xts) # 시계열 패키지
library(caret) # 데이터 표준화 패키지

oil<-getSymbols("DCOILWTICO", src="FRED", auto.assign=FALSE)
US_10Y<-getSymbols("DHHNGSP", src="FRED", auto.assign=FALSE)
US_Breakeven<-getSymbols("DHHNGSP", src="FRED", auto.assign=FALSE)
#FRED에서 WTI, 미국 10년 금리, 미국 BEI(Breakeven Inflation rate) 크롤링
xts<-merge(oil, US_10Y, US_Breakeven) # 위 3개 데이터를 모두 xts에 저장
xts<-xts[index(xts) >"2003-01-04"] # 가장 짧은 시계열에 시작점을 맞춤
xts<-cbind(na.locf(xts)) # 결측치가 있을 경우 그 앞의 데이터를 사용
preproc<-preprocess(coredata(xts), method=c("range"))
norm<-predict(preproc, coredata(xts))
# 시계열에서 인덱스를 제외한 코어를 가져와서 (0,1)사이 값으로 맵핑
```

자료: R, Fred Economic Research, 신한금융투자

클러스터링 R 코드

```
library(quantmod) # 데이터 풀링 패키지
library(xts) # 시계열 패키지
library(factoextra) # 클러스터링 패키지

fviz_nbclust(norm, kmeans, method="wss") # 표준화된 시계열을 사용해
kmeans 클러스터링을 적용할 때 최적 군집수(k)를 찾기 위해 오차값을 도식
km.res = kmeans(norm, 4, nstart = 1000) # 4개 군집으로 나누는 kmeans 알고
리즘을 1,000회 반복하고 그 중 분산값이 가장 작은 것을 사용
fviz_cluster(km.res, data=norm) # 클러스터링 결과 시각화
```

자료: R, Fred Economic Research, 신한금융투자

사전적으로 변수들을 취사선택했어도 변수 개수가 여전히 많기 때문에 추가적인 변수축소가 필요하다. 머신러닝에 근거한 일반적인 차원 축소 방법 중 하나는 주성분분석(PCA:Principal Component Analysis)으로 공통의 설명축을 찾거나 제약회귀를 사용해 투입변수의 차원을 축소하는 것이다. 투입가능 변수 풀의 관측 주기가 변수마다 다르기 때문에 일간/주간/월간으로 공통된 시점에 발표되는 지표끼리 묶어 변수축소를 진행했다.

CORRPLOT, PCA 구현 및 시각화 R 코드

```
library(corrplot) # 상관계수 패키지
library(FactoMineR) # PCA 패키지
# 위에서 데이터를 받아둔 xts 변수를 사용할 것임. 본문에 구현된 부분은 유
료데이터인 Refinitiv를 사용해 엑셀로 데이터를 정리한 뒤 읽어서 사용했음

Corr_mat=cor(xts) # 공분산 행렬 구하기
corrplot=(Corr_mat, method="color", outline=T, addgrid.col="darkgray",
order = "hclust", addrect=4) # CORRPLOT 도식하기
principal_Comp = PCA(xts,graph=FALSE) # PCA 수행
fviz_screeplot(Principal_Comp, addlabels=TRUE, ylim=c(0,50))
# PCA 변환축 설명력 시각화(Screeplot)
fviz_pca_var(principal_Comp, repel=TRUE)
# PCA 결과 투입변수별 변환축 위 방향 시각화
```

자료: Refinitiv, R, 신한금융투자

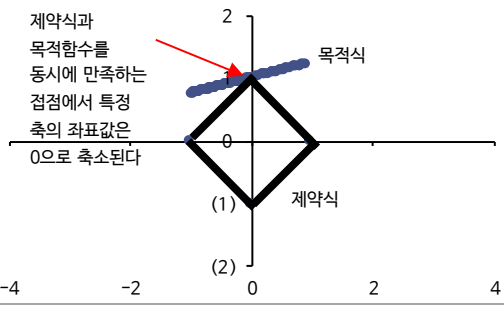
LASSO 회귀 구현 및 시각화 R 코드

```
library(glmnet) # Ridge, LASSO 등 정규화 회귀 패키지
library(vip) # LASSO 결과 주요 변수 도식화
# 위에서 데이터를 받아둔 xts 변수를 독립변수로 사용할 것이며 설명변수로
는 다음달 선진국-신흥국 주가 수익률을 사용. 설명변수 열의 이름은 Result
로 xts 변수 오른쪽 끝에 삽입하면 됨
x=model.matrix(Result~.,xts) # 독립변수
y=xts$Result # 종속변수
glm_Lasso=cv.glmnet(x,y,alpha=1,family="Gaussian")
 # cv:cross validation 교차 타당성 검증
plot(glm_Lasso) # 람다와 MSE 간의 trade-off 도식
fit1= glmnet(x,y,alpha=1,family="Gaussian")
 # alpha=1-> LASSO, alpha=0 -> Ridge, family 조정으로 binomial도 가능
plot(fit1, xvar="lambda") # 람다가 커질수록 변수 축소는 더 잘 기능
vip(glm_Lasso, num_features=10, geom="point") # 중요도 순으로 변수 도식
```

자료: Refinitiv, R, 신한금융투자

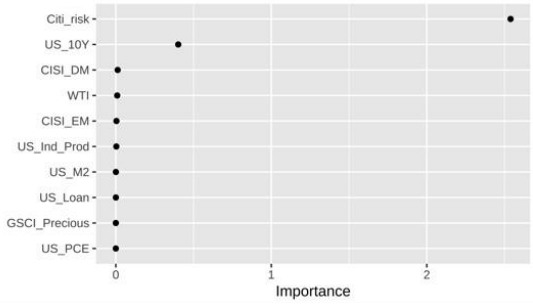
차원축소를 하기 위한 또 다른 방법은 제약회귀의 특수한 형태인 라쏘회귀(LASSO:Least Absolute Shrinkage and Selection Operator)를 사용하는 것이다. 제약조건하 최적화 기법은 목적함수와 제약함수가 접하는 부분에서 해를 찾는다. 이를 이용해 아래와 같이 제약식을 축에 꼭지점을 갖는 사각형으로 설정함으로써 목적함수와 제약함수가 만나는 점이 축 위의 꼭지점이 되도록 하는 것이다.

LASSO 회귀의 변수축소 원리 시각화



자료: R, 신한금융투자

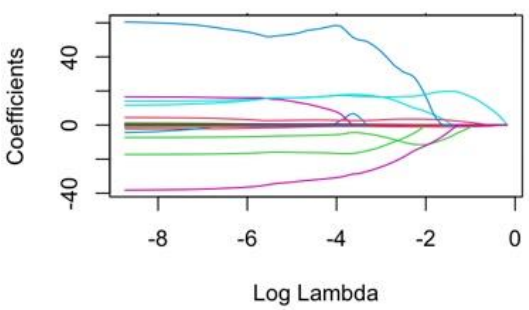
매크로 변수를 투입한 실증분석 결과(vip함수)



자료: Refinitiv, R, 신한금융투자

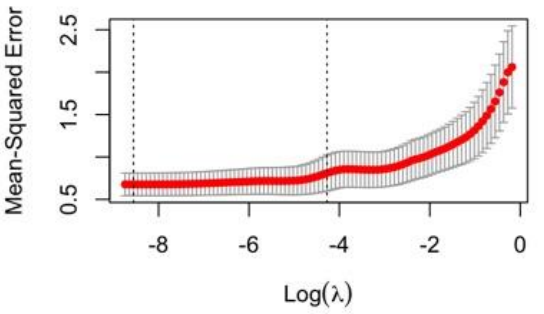
축 위의 꼭지점의 좌표는 (x,0,0 ,⋯,0)과 같은 형태이므로 최적점에서 불필요한 변수의 값은 0이 된다. 즉, 불필요한 변수는 아예 모델에서 제외해버릴 수 있는 기능을 갖는다. 결과를 해석할때 PCA의 경우 특정 변수들의 조합을 새로운 축으로 사용하기 때문에 직관적으로 이해하거나 설명하기 어려워지는 것에 비교해서 불필요 변수를 아예 배제할 수 있다는 점은 다른 기법에 비해 LASSO 기법의 장점이다.

람다를 조정해 변수축소 강도를 조절



자료: Refinitiv, R, 신한금융투자

변수축소와 MSE 사이의 trade-off



자료: Refinitiv, R, 신한금융투자

R에서 라쏘를 구현하는 것 또한 패키지를 사용하면 몇 줄로 구현이 가능하다. 라쏘를 사용하고자 한다면 패널티항의 계수인 람다값을 어느정도로 설정해야할지에 대한 고민이 필요하다. 위 그림과 같이 람다를 높일수록 효과적인 변수축소가 가능하지만 투입변수가 줄어들수록 MSE(Mean-Squared Error)가 늘어나기 때문에 최적의 람다값을 찾기 위해 교차 타당성 검증을 해야한다.

Compliance Notice

- 이 자료에 게재된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다(작성자: 박우열).
- 본 자료는 과거의 자료를 기초로 한 투자참고 자료로서, 향후 주가 움직임은 과거의 패턴과 다를 수 있습니다.
- 자료 제공일 현재 당사는 상기 회사가 발행한 주식을 1% 이상 보유하고 있지 않습니다.
- 자료 제공일 현재 당사는 지난 1년간 상기 회사의 최초 증권시장 상장시 대표 주관사로 참여한 적이 없습니다.
- 자료제공일 현재 조사분석 담당자는 상기회사가 발행한 주식 및 주식관련사채에 대하여 규정상 고지하여야 할 재산적 이해관계가 없으며, 추천의견을 제시함에 있어 어떠한 금전적 보상과도 연계되어 있지 않습니다.
- 당자료는 상기 회사 및 상기회사의 유가증권에 대한 조사분석담당자의 의견을 정확히 반영하고 있으나 이는 자료제공일 현재 시점에서의 의견 및 추정치로서 실적치와 오차가 발생할 수 있으며, 투자를 유도할 목적이 아니라 투자자의 투자판단에 참고가 되는 정보제공을 목적으로 하고 있습니다. 따라서 종목의 선택이나 투자의 최종결정은 투자자 자신의 판단으로 하시기 바랍니다.
- 본 조사분석자료는 당사 고객에 한하여 배포되는 자료로 어떠한 경우에도 당사의 허락없이 복사, 대여, 재배포될 수 없습니다.