비유동적 시장에서의 계량적 투자전략 제안

Quantitative investment under illiquid market

: proposal of the global minimum variance-optimized portfolio with machine learning-based stock direction prediction

Yeongjae On[1] , Giwoon Yoo[1], Minjun Sohn[1], and Jongwoo Lee[1] (2017112535, 2017112533, 2017112549, 2017112576)

[1] Department of Industrial and Systems Engineering, Dongguk University-Seoul, Seoul, Korea

Abstract

코로나19로 인해 찾아온 2020년의 대봉쇄 (The Great Lockdown)로 인한 경제 위기는, 전 세계 주식 시장의 대 폭락을 야기했다. 이에 종목 간 변동성은 매우 불안정해졌는데, 여기에 미국 정부의 이른바 "달러 찍어내기"가 심하게 낮아진 주가와 맞물려 전 세계 지수의 유래 없는 대상승이 발생하게 되었다. 그러나, 최근 미 정부가 테이퍼링을 발표하였고, 여기에 얼마 지나지 않아 한국은행 역시 지난 8월 1차 기준금리 인상 이후 연말에 추가적 인상을 예고한 상황에서 환율 또한 상승세를 보이고 있는 상황이다. 이로 인해 국내 주식시장이 소위 말하는 횡보를 비롯한 하락을 맞이하게 되었고, 일반 투자자들은 현재 심각하게 위축된 상황이다. 현 상황처럼, 자산을 파는 것이 어려워지며 투자자들의 투자 의욕이 떨어진 시장을 비유동적 시장(illiquid market)이라고 한다. 본 연구에서는 이러한 비유동적 시장에서 보다 안정적으로 꾸준한 수익률을 얻을 수 있는 투자 전략을, 계량적 투자 (Quantitative investment)의 관점에서 설계하고자 한다. 우리는 KOSPI 시가총액 150위, KOSDAQ 시가총액 50위에 드는 주식 종목을 대상으로 향후 30영업일 간의 방향성을 예측하는 기계학습 모델을 제작, 그 후 예측한 결과를 바탕으로 분산(위험) 최적화(Global Minimum Variance) 포트폴리오를 제안한다.

Data Preparation

시가총액 기준 KOSPI 150위, KOSDAQ 50위 이내 종목의 일별 종가 → 종속 변수

[1] 거시경제변수, 시장지수, 선물, 환율 등 총 28가지의 일별 종가 → 독립 변수

데이터 제공: [2] FRED, Yahoo, Naver, [3] KRX

[1] 하이일드 채권 스프레드, S&P500, 다우존스 산업지수, 필라델피아 반도체지수, CBOE 변동성 지수, 미국채 10년물, 미국채 5년물, 닛케이225, 항셍지수, 대만 가권, S&P500 선물, 다우존스 산업지수 선물, 나스닥100 선물, 미국채 선물, WTI 유가 선물, 천연가스 선물, 금 선물, 백금 선물, 구리 선물, 미국 달러 지수, 원-달러 환율, 원-엔화 환율, MSCI World, MSCI ACWI, MSCI EM, 국채 10년물, 국채 10년 선물, KOSPI200 변동성 지수

[2] Federal Reserve Economic Data; 미연방준비은행 경제 데이터

[3] Korea Exchange; 한국거래소

	ICE BofA US High Yield Option-Adjusted Spread	S&P500	Dow Jones Industrial Average Index	NASDAQ
1999-12-31	NaN	NaN	NaN	NaN
2000-01-03	NaN	1469.250000	11497.120117	4069.310059
2000-01-04	4.68	1455.219971	11357.509766	4131.149902
2000-01-05	4.81	1399.420044	10997.929688	3901.689941
2000-01-06	4.77	1402.109985	11122.650391	3877.540039
...	...	...	...	...

K-means Clustering

각 종목의 일별 종가 데이터 수집

↓

2020년 1월 이후의 데이터만 사용[1]

↓

최근 상장한 종목의 경우 관측치 개수의 문제로 제외

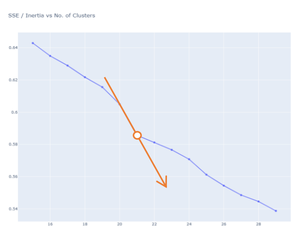
↓

일별 종가를 일별 수익률로 차분(Differencing)

↓

일별 수익률 데이터 정규화(Normalize)

K의 값을 15에서 29까지 바꿔가면서, 각 K의 오차제곱합(SSE)을 그래프화



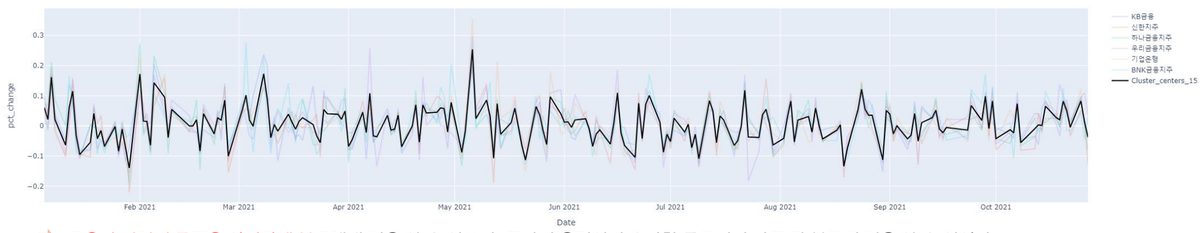
✓ SSE가 급격하게 감소하는 구간 선정 (Elbow method)

✓ K=21일 때가 최적의 군집

✓ Euclidean 거리를 척도로 사용

[1] 최근 시장상황의 변동성을 최대한 반영하기 위해, 코로나19로 인한 대봉쇄 이후의 데이터를 사용함

K-Mean Clustering Algorithm을 사용해 군집화 된 결과 (Cluster 15)



금융과 관련된 종목을 완벽하게 분류해낸 것을 알 수 있으며, 주가의 움직임이 유사한 종목끼리 아주 잘 분류된 것을 알 수 있었다.

마코위츠의 포트폴리오 이론에 따른 분산투자를 실행할 시, **유사한 움직임을 보이는 종목을 겹쳐서 선정하는 것을 방지**하여 종목을 구성하게끔 한다.

Data Analysis & Granger Causality Test

추후 진행할 Random Forest는 블랙박스(Black Box) 모델
∴ 미리 인과관계가 존재하지 않는 변수를 제외시켜 모델의 신뢰성과 해석력을 높일 필요성이 존재

Granger Causality Test 진행
각 종목(189개) 기준 30영업일이 지난 시점에도 각 변수(28개)가 그레인저 인과관계가 있는지, 유의 수준 0.05로 검정

↓

정상적인 시계열 데이터의 분석을 위하여 변수, 종목의 종가 데이터의 차분(Differencing) 진행

종목 데이터 총 200개, 일별, 30영업일별 수익률로 차분

변수 데이터 총 28개, 일별 수익률로 차분

↓

통계적으로 시계열이 정상성을 나타내는지 확인하기 위해 [1] **Augmented Dickey-Fuller Test**를 사용, 유의수준 0.05로 검정

[1] Granger Causality Test에서는 입력된 시계열이 모두 정상성을 나타낸다는 전제 하에 진행되므로, ADF Test를 통한 정상성 검정이 필수적으로 수행되어야 함

↓

일별: 차분 결과 모두 정상성을 나타냄

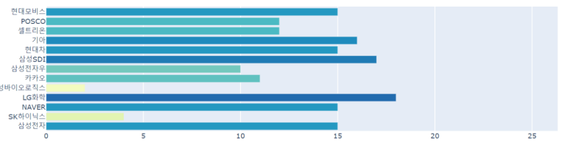
30영업일별: 차분 결과 총 11종목이 유의하지 않음(제외)

일별: 차분 결과 모두 정상성을 나타냄

```
from statsmodels.tsa.stattools import grangercausalitytests

def granger_result_df(stock, features, lag):
    table = pd.DataFrame(columns = ['Name', 'Granger Causality'])
    try:
        for col in features.columns.tolist():
            df = pd.concat([stock, features[col]], axis=1).dropna()
            granger = grangercausalitytests(df.values, maxlag=[lag], verbose=False)
```

Features	삼성전자	SK하이닉스	NAVER	LG화학	삼성바이오로직스	카카오	삼성전자우	삼성SDI	현대차	기아	...	한화에어로스페이스	휠라홀딩스	효성	SK케미칼	롯데정밀화학	영원무역
ICE BofA US High Yield Option-Adjusted Spread	True	True	True	False	False	False	True	True	True	True	...	True	True	True	True	True	True
S&P500	True	True	True	True	False	True	True	True	True	True	...	True	True	True	False	True	True
Dow Jones Industrial Average Index	True	True	True	True	False	True	True	True	True	True	...	True	True	True	False	True	True



Granger Causality Test에 대한 결과 해석

H_0: 변수는 종목의 30영업일 종가에 대하여 Granger Causality를 따르지 않는다.
H_1: 변수는 종목의 30영업일 종가에 대하여 Granger Causality를 따른다.
True: 귀무가설을 기각한다.
False: 귀무가설을 기각하지 못한다.

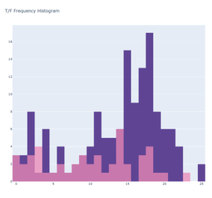
T/F 빈도분석

KOSPI T/F Analysis

	True	False
count	140.000000	140.000000
mean	13.528571	14.271429
std	6.162596	6.174726
min	0.000000	0.000000
25%	10.000000	10.000000
50%	15.000000	13.000000
75%	18.000000	17.250000
max	25.000000	28.000000

KOSDAQ T/F Analysis

	True	False
count	48.000000	48.000000
mean	10.020833	17.979167
std	6.336284	6.336284
min	0.000000	5.000000
25%	3.750000	13.750000
50%	10.500000	17.500000
75%	14.250000	24.250000
max	23.000000	28.000000



KOSPI의 Q1 = KOSDAQ의 Mean = 10개
Granger Causality Test를 통과한 변수의 수가 10개 미만인 종목은 추후 학습에 제외

처음 시작한 200개의 종목 중 134개 선정

Stock Direction Prediction using Randon Forest

종목의 일별 종가 데이터를, **"30영업일이 지났을 때 매도했을 경우의 수익률(R_{30})"**로 변환

R_{30}이 4% 초과면 "Up", 0%이상 4%이하면 "Neutral", 0%미만이면 "Down"으로 Class를 분류한 후 학습 => **예측 결과는 "Up", "Neutral", "Down" 총 3종류**

$$R_{30} = \frac{30\ business\ day\ after's\ Close\ - Today's\ Close}{Today's\ Close}$$

```
def direction_label(Series):
    if Series > 0.04: return 'Up'
    elif Series>=0 and Series <= 0.04: return 'Neutral'
    else: return 'Down'
```

모델 설계 (총 134개)

Train-validation data split method: Stratified Shuffle Split, Test size = 30%; 데이터 누설(Data Leakage)과 클래스 불균형을 해소

Hyper parameter tuning: RandomizedSearchCV; Stratified 10-fold cross validation를 진행한 후 Best parameter 선정

****Parameter distribution:** n_estimators (50, 100, 150), max_depth (20, 30, 40, 50), max_features (n, n-2, n-4, n-6)

**n = number of features

Random Forest Classifier: RandomizedSearchCV에서 찾은 Best parameter을 통하여 학습

Evaluation method: Stratified 10-fold cross validation score, Confusion matrix

```
sss = StratifiedShuffleSplit(n_splits=1, test_size=0.3, random_state=0)
rnd_clf = RandomForestClassifier(n_estimators=100, n_jobs=-1, random_state=0)
rnd_search = RandomizedSearchCV(rnd_clf, param_dist, cv=10, random_state=0)
```

```
rnd_clf = RandomForestClassifier(**param_best, n_jobs=-1, random_state=0)
rnd_scores = cross_val_score(rnd_clf, X_train, y_train, scoring='accuracy', cv=10)
y_test_pred = rnd_clf.predict(X_test)
cm_test = confusion_matrix(y_test, y_test_pred, labels=['Up', 'Neutral', 'Down'])
```

Model Results

```
삼성전자 Direction Prediction Model
Best Parameter: {'n_estimators': 100, 'max_features': 9, 'max_depth': 20}
NAVER Direction Prediction Model
Best Parameter: {'n_estimators': 100, 'max_features': 9, 'max_depth': 40}
LG화학 Direction Prediction Model
Best Parameter: {'n_estimators': 100, 'max_features': 12, 'max_depth': 20}
10-fold Cross Validation, Accuracy score mean: 80.432%
Accuracy score of Test Data(30% of sample): 80.763%
[Confusion Matrix of Test Data]
      Up Neutral Down
[[237   9  20]
 [ 29  26  33]
 [ 21  14 266]]
--Model Successfully Saved--
```

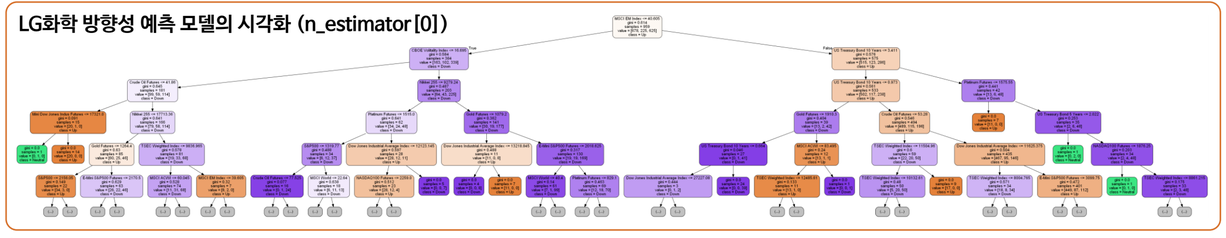
각 종목별로 선정된 최적 파라미터

Cross Validation Score

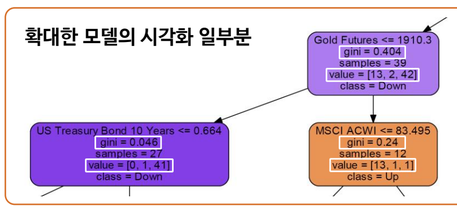
Validation Data에 대한 정확도

Confusion Matrix; Validation Data의 정답 Class를 얼마나 잘 예측했는지 직접 확인

LG화학 방향성 예측 모델의 시각화 (n_estimator[0])

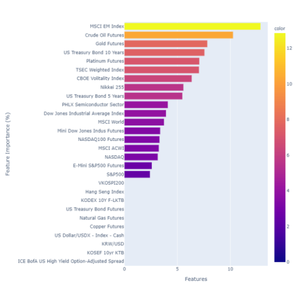


확대한 모델의 시각화 일부분



노드가 분리되면서, Gini Impurity Index가 감소함과 동시에 자료가 더 잘 나누어짐을 확인할 수 있다.

	ICE BofA US High Yield Option-Adjusted Spread	S&P500	Dow Jones Industrial Average Index	NASDAQ	PHLX Semiconductor Sector	CBOE Volitality Index
삼성전자	0.085283	0.034411	0.038993	0.053656	0.059485	0.087677
NAVER	0.101256	0.035796	0.041801	0.042800	0.072889	0.000000
LG화학	0.000000	0.024226	0.039355	0.031590	0.041073	0.063631
카카오	0.000000	0.056412	0.062447	0.083641	0.000000	0.089553
삼성전우	0.130718	0.062073	0.078740	0.000000	0.109357	0.000000
...	...	...	...	...	...	...
에지온	0.097788	0.052037	0.054333	0.000000	0.000000	0.000000
바이오니아	0.093194	0.023929	0.025254	0.030954	0.041179	0.064071
이오테크닉스	0.074270	0.027369	0.025241	0.028059	0.051573	0.062345
고영	0.091852	0.033095	0.052613	0.035410	0.050118	0.063905
에스에프에이	0.087266	0.038761	0.045043	0.057014	0.065365	0.000000



전 종목에 대하여 Feature Importance를 확인할 수 있게끔 데이터프레임의 형태로 제작하고, 이를 시각화하여 블랙박스 모델의 최대 단점인 신뢰성과 해석력을 보완하고자 하였다.

Back Testing with Random Forest Model

주가의 움직임을 예측한 경우, 모델의 성능을 단순히 metric으로만 평가하면 모델을 과대평가할 가능성이 생김

↓

모델이 예측한 결과를 이용하여, 실제 과거 데이터에 대한 투자 시뮬레이션을 진행하여, 모델을 바탕으로 투자하는 것을 검증

투자 알고리즘 예시

[시작일]의 Class : Up(무조건 Up 고정) => Buy(매수)
[시작일 이후 30영업일]의 Class: Up => Hold
[시작일 이후 60영업일]의 Class: Neutral => Hold
[시작일 이후 90영업일]의 Class: Down => Sell(매도)
[시작일 이후 120영업일]의 Class: Down => Stay(=Hold, 중지)
[시작일 이후 150영업일]의 Class: Up => Buy(다시 매수)
(..반복..)

2019-09-25	Buy (up sign)	305000	NaN	NaN
2019-11-20	Sell (down sign)	309500	1.014754	18.954404
2020-01-15	Buy (up sign)	328500	NaN	NaN
2020-03-12	Hold (neutral sign)	NaN	NaN	NaN
2020-05-14	Hold (up sign)	NaN	NaN	NaN
2020-07-14	Hold (up sign)	NaN	NaN	NaN
2020-09-09	Sell (down sign)	699000	2.127854	41.460057
2020-11-09	Buy (up sign)	734000	NaN	NaN
2020-12-23	Hold (up sign)	NaN	NaN	NaN
2021-03-02	Sell (down sign)	891000	1.213896	50.542113

LG화학 모델에 대한 투자 시뮬레이션

결과, 상당한 양의 수익률을 확인할 수 있음

=> 모델을 바탕으로 30영업일 간격으로 투자하는 방법을 고려해 볼 법만 함

```
Trading CAGR: 39.537%
Trading Annual Standard deviation: 26.227%
Trading Sharpe Ratio: 1.507
```

Global Minimum Variance Portfolio Optimization

많이 알려져 있는 Mean-variance optimization은 목적함수가 **Sharpe Ratio를 최대화하며**, 최적화에 각 종목의 기대수익률을 요구하는데, 하락 추세의 시장에선 기대수익률 예측이 거의 불가능해 오히려 좋지 못한 결과를 초래할 수 있음

↓

모든 자산의 기대수익률이 동등하다고 가정하며, 목적함수가 **Variance를 최소화하는, Global minimum variance (GMV; 전역최소분산)**는 오히려 시장이 하락 추세일 때 다른 포트폴리오 최적화 방법보다 안정되고 우수한 결과를 보여줌

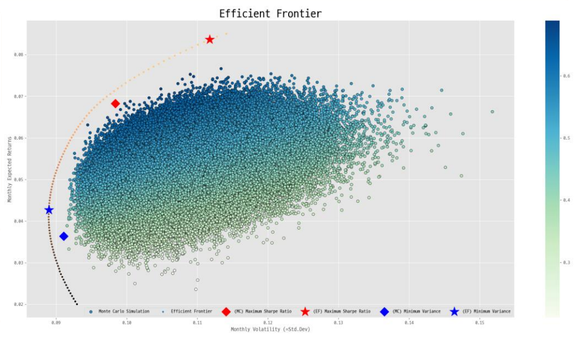
학습된 모든 모델에 현 시점 (2021/10/26)의 변수 데이터를 입력하여 **현 시점에서의 방향성이 "Up"인 종목**을 골라낸 후, **K-means Clustering으로 분류된 군집에 대하여 동일 군집에 속하지 않은 종목을 선정**하여 포트폴리오 구성

↓

선정 종목(군집): OCI(2), 한화솔루션(6), 한국가스공사(7), 일진머티리얼즈(8), 한솔케미칼(11), 메리츠금융지주(12), 이마트(13), KB금융(15), 기업은행(15), LG화학(18)

↓

총 30만회의 몬테카를로 시뮬레이션의 포트폴리오의 가능한 분포를 확인, 그 후 **SLSQP (Sequential Least Squares Programming) 알고리즘을 이용하여 효율적 프론티어와 GMV 포트폴리오의 [1] 최적해**를 구함



[1] 계산에 필요한 무위험자산수익률은 국고채 3년 수익률을 월 수익률로 환산하여 사용함

	OCI	한화솔루션	한국가스공사	일진머티리얼즈	한솔케미칼	메리츠금융지주	이마트	KB금융	기업은행	LG화학
(EF) Minimum Variance	0.000	0.000	5.509	4.257	11.146	16.482	34.827	0.000	21.216	6.563

Conclusion & Limitation

✓ 완성된 최적 GMV 포트폴리오는 월기대수익률 4.268%, 월변동성 8.897%, 샤프비율 0.480, 베타 0.207을 보여준다. 이는 시장(KOSPI 1배 추종)의 월기대수익률 2.356%, 월변동성 8.160%, 샤프비율 0.284임을 고려하면, 비슷한 수준의 변동성을 갖고 있으나 약 1.8배 가량 더 높은 수익률을 보여준다고 해석할 수 있다. 특히, 베타는 0.207로 이는 매우 낮은 수준인데, 완성된 포트폴리오는 시장이 1% 변동할 때 0.207%만 변동한다고 해석이 가능하다. 즉, 이는 초기 목적이었던 "비유동적 시장에서 사용가능한 투자전략 제안"에 매우 어울리는 결과라고 볼 수 있다.

✓ 완성된 포트폴리오의 베타계수는 매우 낮다는 점은 오히려 시장이 상승세를 띄고 있는 경우엔 좋은 성과를 내기가 어렵다. 또한, 위험 회피적 성향이 매우 강한 포트폴리오이기에 높은 위험 프리미엄을 원하는 투자자에게는 적합하지 않다.

✓ 데이터의 기간이 동일하지 않아 일관성 있는 모델 제작이 불가능했으며, 몇몇은 그 기간이 매우 짧아 모델 제작 자체가 불가능하였다.

✓ 주가에 영향을 주는 독립변수 선택에 있어 한정된 변수만을 선택했다. 소비자 물가지수, 소비자 심리지수 등 더욱 시장 움직임을 잘 설명하는 변수들은 다수 존재하지만, 일반 투자자에겐 월별 또는 분기별로 제공되는 등 접근하는 데에 한계가 있어 사용하지 못했다. 이러한 데이터의 질과 양의 차이로 인해 모델의 해석능력에 아쉬움이 남는다.