

# 포트폴리오 복제하기

## 펀드는 어떤 종목에 투자하고 있을까?

### 포트폴리오 추론 문제

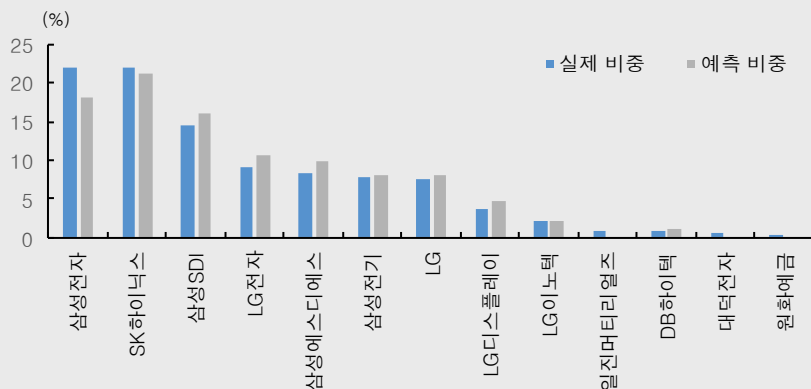
어떤 펀드의 기준가 추이를 알고 있다고 할 때, 우리는 이 펀드가 어떤 종목에 투자하고 있는지 알 수 있을까? 포트폴리오 추론 문제를 풀기 위해서 먼저 포트폴리오가 어떤 종목을 담고 있는지 추정해야 한다. 이를 위해 변수선택 알고리즘을 사용했다. 변수선택은 변수들의 조합 가운데 가장 적합한 변수만을 추출하는 과정이다. 순차적 변수선택 알고리즘을 활용해 펀드의 구성종목을 추정하고, 회귀분석으로 종목의 비중을 계산했다.

실제 국내주식 섹터 ETF를 복제하는 포트폴리오를 구성해 보았다. ETF와 같이 패시브하게 운용되며 구성 종목의 변화가 많지 않은 경우에 포트폴리오 편입 종목과 비중을 효과적으로 예측한다. 국내 주식 ETF를 대상으로 할 때 90% 이상의 높은 적중률을 나타냈다.

### 포트폴리오 복제를 통한 트레이딩 활용

알고리즘을 활용한 포트폴리오 추론은 제한된 정보 하에서 효과적으로 포트폴리오를 복제하는 방법이다. 트레이딩 기회를 확보하거나 펀드 매니저의 운용 행태를 확인할 수 있다는 점에서 유용하다고 판단한다. 목표 포트폴리오의 지수나 펀드의 가격(기준가)이 공시되는 투자대상이면 가능하다. 또한 펀드가 유동성이 낮은 자산으로 운용이 되고 있다면 동일한 방법을 거쳐 유동성이 높은 자산으로 복제하는 것도 검토할 수 있다.

### TIGER200 IT ETF의 실제와 예측 비중



자료: 미래에셋자산운용, 한국투자증권

### 목차

I. 포트폴리오는 어디에 투자하고 있을까?	1
포트폴리오 추론 문제	
변수선택 알고리즘	
II. 케이스 스터디	3
포트폴리오 편입종목 추정	
포트폴리오 비중 추정	
III. 투자 시사점	9
포트폴리오 복제를 통한 트레이딩 활용	

정현종, CFA

hyeonjong.jung@truefriend.com

윤지수

jisu.yoon@truefriend.com

# I. 포트폴리오는 어디에 투자하고 있을까?

## 포트폴리오 추론 문제

펀드의 기준가 정보만 가지고 매니저가 투자하고 있는 종목과 비중을 추정할 수 있을까?

어떤 펀드의 기준가 추이를 알고 있다고 할 때, 우리는 이 펀드가 어떤 종목에 투자하고 있는지 알 수 있을까? 모든 경우의 수를 검토하는 것은 불가능하다. KOSPI200 지수 구성종목을 유니버스로 하는 펀드가 30종목에 투자하고 있다면 가능한 조합은  ${}_{200}C_{30}$ 으로  $4.1 \times 10^{35}$ 개에 달하기 때문이다.

여기서는 변수선택 알고리즘을 활용해 펀드의 구성종목을 추정하고, 회귀분석으로 종목의 비중을 계산했다. 실제 국내주식 섹터 ETF를 복제하는 포트폴리오를 추론해 보았다. ETF와 같이 패시브하게 운용되며 구성 종목의 변화가 많지 않은 경우에 포트폴리오 구성종목과 비중을 효과적으로 예측하는 것으로 나타난다.

펀드가 룡쫓 전략을 쓰거나 쫓은 리밸런싱을 거치는 경우, 기준가 확인이 힘든 경우에 성과를 트래킹하기 어려울 수 있다. 그러나 알고리즘을 활용한 포트폴리오 추론은 제한된 정보 하에서 효과적으로 포트폴리오를 복제할 수 있는 방법으로 트레이딩 기회를 확보하거나 펀드 매니저의 운용 행태를 확인할 수 있다는 점에서 유용하다고 판단한다.

## 변수선택 알고리즘

포트폴리오 추론 문제, 많은 변수 가운데 어떻게 효과적인 변수만을 선택할까?

포트폴리오 추론 문제를 효율적으로 풀기 위해서 먼저 포트폴리오가 어떤 종목을 담고 있는지 추정해야 한다. 이를 위해 변수선택 알고리즘을 사용했다. 변수선택은 변수들의 조합 가운데 가장 적합한 변수만을 추출하는 과정이다. 변수선택 알고리즘을 활용해 자동으로 최적의 변수 조합을 추려낼 수 있다면 연산 시간을 줄이고 입력 변수 가운데 관계가 없는 노이즈를 제거할 수 있을 것이다.

일반적으로 변수의 차원을 축소하기 위해 주성분 분석을 쓰거나 Lasso 회귀분석을 통해 일반화(regularization)하는 방법이 쓰이는데, 여기서는 순차적 변수선택 알고리즘을 사용한다. 순차적 변수선택 알고리즘은 모든 경우의 수를 계산하기가 불가능할 때 선택할 수 있다. 탐욕적 검색 알고리즘(Greedy search algorithm)의 일종으로 관련성이 낮은 변수를 제거해 전체 변수의 차원을 축소하는 방법이다. 원하는 변수의 개수에 도달할 때까지 순차적으로 하나씩 변수를 추가하거나 제거하는 과정을 거치게 된다<해설도우미 참조>.

여기서 변수를 선택하는 기준은 MSE(Mean Squared Error)를 사용했다.  $n$ 은 샘플의 크기,  $y_i$ 는 실제값,  $\hat{y}_i$ 는 예측값을 각각 의미한다. 실제값(목표 포트폴리오의 수익률)과 예측값(모델로 예측한 수익률) 간의 차이를 의미하는 평균 제곱오차가 최소화되도록 변수를 선택하는 과정이라 할 수 있다.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

## <해설도우미> 순차적 변수선택 알고리즘

순차적 변수선택 알고리즘은 다음 4가지로 나뉜다.

1. Sequential Forward Selection(SFS)
2. Sequential Backward Selection(SBS)
3. Sequential Floating Forward Selection(SFFS)
4. Sequential Floating Backward Selection(SFBS)

SFS나 SBS는 기본적인 변수선택 알고리즘인데, 선택된 모델의 성과가 개선되는 경우에만 특정 변수를 추가하거나 제외한다. SFS는 목표 포트폴리오의 성과를 가장 잘 예측하는 1개의 변수에서 시작해 이후 추가적인 개선이 일어날 때마다 변수를 더해 나가는 방식이다.

SBS는 반대로 모든 변수를 포함한 모델에서 시작해 변수를 하나씩 제거해 추가적인 개선이 이뤄지도록 진행된다. 결국 매우 많은 변수가 존재할 때, 결과값을 잘 설명하는 변수의 부분 조합을 고르는 알고리즘이다.

Floating 알고리즘은 SFS나 SBS를 약간 변형한 것인데 더 다양한 변수들간의 조합을 검토한다. SFFS는 SFS와 같이 변수의 개수를 늘려나가는 방식이지만, 변수의 추가와 제거가 번갈아 나타날 수 있다는 점에서 다르다.

SFFS 알고리즘은 다음과 같은 단계로 진행된다.  $d$ 는 전체 변수의 개수를,  $k$ 는 선택된 변수 조합에서 변수의 개수를 각각 의미한다( $k < d$ ).

### 초기 시점

초기 시점에 알고리즘은  $k=0$ 에서 시작

### 1단계(추가)

선택 기준에 따라 결과값을 가장 잘 설명하는 변수를 추가. 2단계로 이동

### 2단계(조건부 제거)

변수 한 개를 제거했을 때 선택된 변수 조합이 결과값을 더 잘 설명한다면 해당 변수를 제거하는 과정을 반복. 그렇지 않으면 1단계로 이동

### 종료 시점

종료 기준에 도달할 때까지 1단계와 2단계를 반복적으로 수행. 선택된 변수의 개수( $k$ )가 사전에 지정한 변수의 개수( $p$ )와 동일할 때 종료( $k = p$ )

## II. 케이스 스터디

### 포트폴리오 편입종목 추정

국내 주식 ETF의 편입  
종목과 비중을 어떻게  
추정할 수 있을까?

펀드의 구성 종목에 대한 아무런 정보없이 기준가와 개별종목 주가 정보만으로 펀드가 어느 종목에 투자하고 있는지 추정해 보자. 추정해야 하는 목표는 해당 펀드의 ‘구성 종목’과 ‘비중’ 두가지이다.

여기서 목표 펀드 또는 목표 포트폴리오는 국내 주식 업종 ETF이다. ETF 운용사는 자산구성 내역인 PDF (Portfolio Deposit File)를 매일 공시하기 때문에 알고리즘에 따른 예측 종목이 실제로 포함되어 있는지 확인하는 것이 가능하다.

목표 포트폴리오로 시가총액이 큰 ETF 가운데 TIGER200 섹터 ETF 4개를 선택했다. 이들 ETF는 KOSPI200 섹터지수를 추적대상지수로 하며 순자산가치(NAV)의 변동률을 기초지수의 변동률과 유사하도록 투자신탁재산을 운용한다. 기초지수 추종을 위해 지수를 구성하는 종목 전체를 편입하는 완전복제전략을 원칙으로, 필요할 경우 최적화 기법을 적용해 부분복제전략을 사용한다.

〈표 1〉 목표 포트폴리오

코드	목표 포트폴리오	AUM(억원)	편입 종목수
A139260	TIGER200 IT ETF	2,570	12
A139230	TIGER200 중공업 ETF	152	11
A139250	TIGER200 에너지화학 ETF	110	25
A139220	TIGER200 건설 ETF	110	14

주: 2019년 10월말 기준  
자료: Quantiwise, 한국투자증권

유니버스는 KOSPI200 지수에 편입된 201개 종목이다. 유니버스는 정기적인 변경이 이뤄진다. 한국거래소는 일평균시가총액 및 일평균 거래대금을 고려해 KOSPI200 지수 정기변경을 1년에 한번(매년 6월) 실시한다.

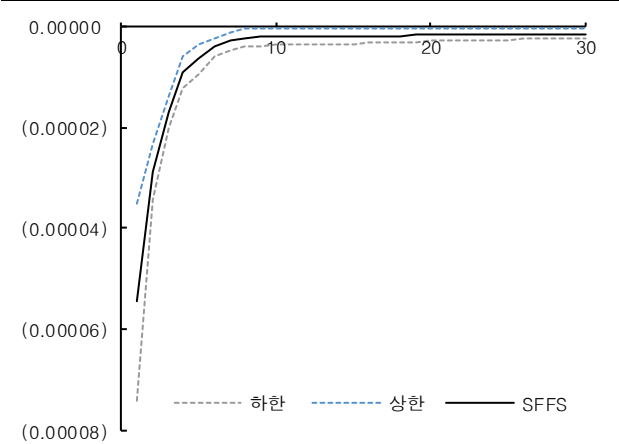
또한 지난해 한국거래소는 GICS 분류체계 개편에 따라 일부 섹터지수의 명칭을 변경하고 구성종목도 교체한 바 있다. 이에 따라 KOSPI200 IT지수를 구성하고 있던 4종목(NAVER, 엔씨소프트, 넷마블, 카카오)이 제외되고 이를 추종하는 섹터 ETF의 편입 종목도 변화했다. GICS 분류체계 변경을 반영해 테스트의 기간은 지난해 12월 14일 이후부터 올해 10월말까지로 하고 일간 데이터를 사용했다.

변수선택 알고리즘(SFFS)에 따라 변수를 선택하고 복제 포트폴리오를 구성했다. 여기서 변수는 유니버스 내의 개별 종목이다. 적합한 변수(종목)만을 조합해 복제 포트폴리오를 구성한 후 목표 포트폴리오와의 수익률 괴리를 계산한다. 복제 포트폴리오의 수익률은 5겹 교차검증(cross-validation)에서 검증 데이터를 사용했을 때의 수익률이다. 5겹 교차검증은 전체 데이터를 5등분해서 5번 모델링을 거쳐 모델을 평가하는 방법이다.

[그림 1~2]에서 가로축은 복제 포트폴리오의 변수(종목) 개수를, 세로축은 목표 포트폴리오와 복제 포트폴리오 수익률 간의 오차(MSE)에  $-1$ 을 곱한 값을 의미한다. 점선은 90% 신뢰구간을 나타낸다.

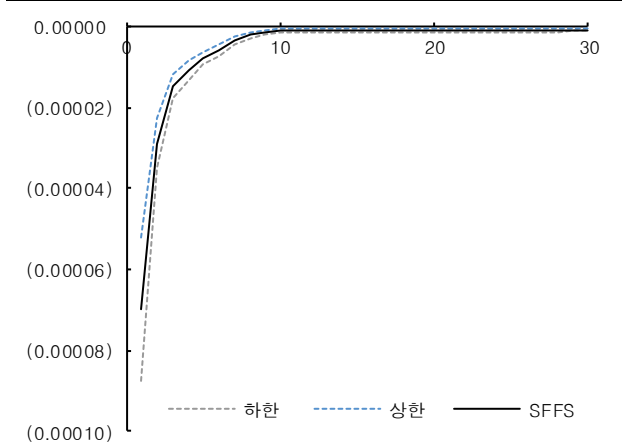
섹터 ETF를 복제하는 과정에서 선택된 변수의 개수가 늘어날수록 수익률 간의 오차는 줄어든다. 또한 10개 내외의 종목만으로 수익률 오차를 대부분 줄일 수 있음을 알 수 있다. 다른 ETF의 결과도 유사하다.

[그림 1] TIGER200 IT ETF 복제포트폴리오의 오차



자료: Quantiwise, 한국투자증권

[그림 2] TIGER200 중공업 ETF 복제포트폴리오의 오차

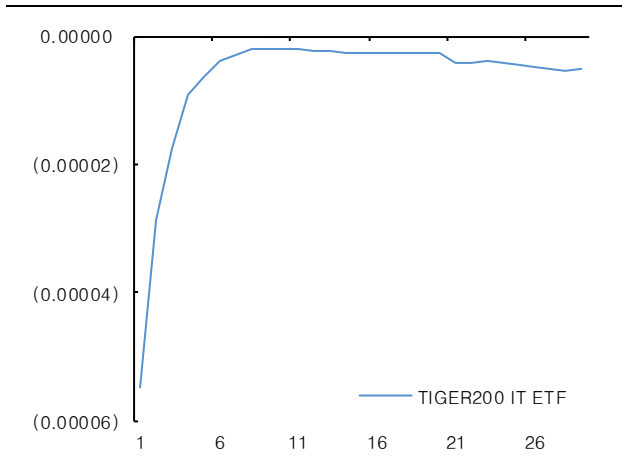


자료: Quantiwise, 한국투자증권

**목표와 복제 포트폴리오간  
수익률 오차 최소화하는  
최적 변수의 개수 추정**

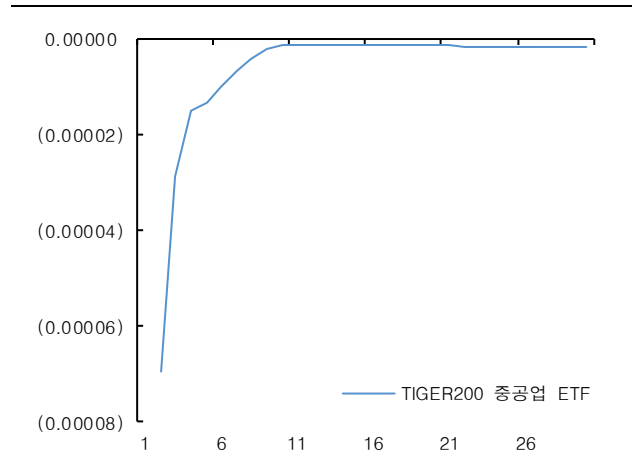
다음으로 그리드서치(Grid search)에 따라 최적 변수의 개수를 산출했다. 그리드서치란 모델의 파라미터 조합을 통해 최적의 모델을 선정하는 방법이다. 변수 선택 알고리즘(SFFS)에서 찾고자 하는 변수의 개수를 1개에서 시작해 늘릴 때마다 수익률의 오차 변화를 확인할 수 있다. TIGER200 IT와 TIGER200 중공업 ETF를 복제하는 포트폴리오에서 오차를 최소화하는 변수의 개수는 모두 10개이다.

[그림 3] 그리드서치와 최적 변수의 개수



자료: Quantiwise, 한국투자증권

[그림 4] 그리드서치와 최적 변수의 개수



자료: Quantiwise, 한국투자증권

변수 선택 알고리즘,  
기준가 정보만으로 목표  
포트폴리오의 실제 편입종목  
추정

변수선택 알고리즘이 얼마나 효과적으로 실제 보유종목을 맞췄는지 비교해 보았다. 섹터 ETF의 실제 종목은 10월말 기준이며, 예측 종목은 지난해 12월부터 올해 10월말까지 ETF 기준가 정보만을 활용해 추정한 종목이다. 섹터 ETF별로 실제 맞춘 종목수는 다음과 같다.

<표 2> 목표 ETF별 실제와 예측 종목수 비교

코드	목표 ETF	AUM(억원)	실제 종목수	예측 종목수	맞춘 종목수
A139260	TIGER200 IT ETF	2,570	12	10	10
A139230	TIGER200 중공업 ETF	152	11	10	10
A139250	TIGER200 에너지화학 ETF	110	25	9	7
A139220	TIGER200 건설 ETF	110	14	12	12

자료: 미래에셋자산운용, 한국투자증권

예측 알고리즘의 정확도를 살펴보기 위해서 혼동행렬을 사용했다. TIGER200 IT ETF의 실제 종목수는 12개인데 예측 알고리즘은 10개를 예측했고, 예측한 종목은 모두 실제 종목이었다(True Positive). 반면 2종목은 예측하지 못했지만 실제 종목인 경우다(False Negative).

여기서는 *Accuracy*와 *MCC(Mattew Correlation Coefficient)*를 통해 예측의 정확도를 확인했다. *Accuracy*는 단순히 전체 중에서 올바르게 예측한 것이 몇 개인지 판단할 때 사용된다. 반면 클래스의 크기가 불균형할 경우, *MCC*가 예측의 정확도를 확인하는데 더 효과적이다. *MCC*는 상관계수로 해석할 수 있는데, 1에 가까울수록 완벽한 예측을 의미한다<해설도우미 참조>.

[그림 5] 혼동행렬(Confusion Matrix) 개념

		Prediction	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

자료: 한국투자증권

[그림 6] TIGER200 IT ETF의 혼동행렬

		Prediction	
		Positive	Negative
Actual	Positive	10	2
	Negative	0	189

자료: Quantwise, 한국투자증권

<표 3> 알고리즘 예측의 정확도

코드	목표 포트폴리오	Accuracy	MCC
A139260	TIGER200 IT ETF	0.99	0.91
A139230	TIGER200 중공업 ETF	0.99	0.95
A139250	TIGER200 에너지화학 ETF	0.90	0.43
A139220	TIGER200 건설 ETF	0.99	0.92

자료: 미래에셋자산운용, 한국투자증권

ETF의 경우, 변수선택  
알고리즘은 높은 정확도로  
종목을 선별

4개 ETF 모두 90% 이상의 *Accuracy*를 보였고, 3개 ETF의 경우 *MCC*가 90%를 상회해 높은 적중률을 나타냈다. 한편 ETF는 추적하는 지수의 변경에 따라 일부 종목이 변화된다. 실제 관찰 기간 중 KOSPI200 정기변경과 섹터지수 변경으로 섹터 ETF의 편입종목 변경이 발생했다. 예를 들어 효성중공업은 기존에 KOSPI200 에너지화학 ETF에 포함되어 있었지만 올해 6월 정기변경에서 중공업 ETF에 새로 편입됐다.

패시브하게 운용되는 ETF 외에 뮤추얼 펀드도 리밸런싱 시점에 구성종목이 변화된다. 이를 동적으로 파악하기 위해서는 관찰 윈도우를 줄여 종목의 변화를 확인할 수 있을 것이다. 예를 들어 매니저의 리밸런싱으로 인한 종목의 편입/제외를 확인하기 위해서 변수선택 알고리즘에 사용되는 기간을 롤링으로 해서 변화를 추적하는 방법을 쓸 수 있다.

### <해설도우미> 혼동행렬과 예측의 정확도

혼동행렬(Confusion Matrix)은 알고리즘의 성능을 평가할 때 많이 사용되는 지표다. True는 실제와 예측이 일치하는 경우를, False는 실제와 예측이 불일치하는 경우를 의미한다. Positive는 있다고 예측한 경우를, Negative는 없다고 예측한 경우를 나타낸다.

따라서 TP(있다고 예측했는데 있는 경우) 또는 TN(없다고 예측했는데 없는 경우)은 정답에 해당한다. 반면 FP(있다고 예측했지만 없는 경우)와 FN(없다고 예측했지만 있는 경우)은 잘못 예측한 경우이다.

혼동행렬로 예측의 정확도를 측정할 수 있는데 다음과 같은 지표를 활용한다. *Accuracy*는 전체 중에서 올바르게 예측한 것이 몇 개인지 판단할 때 사용된다. 그러나 클래스의 크기가 차이가 날 경우, 예측 정확도는 잘못된 정보를 줄 수 있다. 여기에서도 유니버스(KOSPI200)에 속한 종목은 대부분 목표 포트폴리오(섹터 ETF)에 속하지 않는다. 그러나 실제로 해당 ETF에 속한 종목이 하나도 없다고 예측한 경우에도 *Accuracy*는 90% 내외를 기록하게 된다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

반면 *MCC*(*Mattews correlation coefficient*)는 불균형한 클래스의 크기를 감안해 분류의 정확도를 측정한다는 장점이 있다. *MCC*는 상관계수로 해석할 수 있는데, +1일 경우 완벽한 예측을, 0일 경우 우연에 따른 예측을, -1일 경우 잘못된 예측을 각각 의미한다.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP \times FP)(TP + FN)(TN + FP)(TN + FN)}}$$

### 포트폴리오 비중 추정

선별된 종목 비중을 추정해  
목표 펀드와 유사한  
포트폴리오 구성 가능

포트폴리오를 복제하기 위해서는 목표 포트폴리오의 종목을 예상한 후에 비중을 추정해야한다. 테스트 기간 동안 목표 포트폴리오(섹터 ETF)와 추정 종목의 일간 수익률을 회귀 분석해 종목의 편입비를 추정할 수 있다.

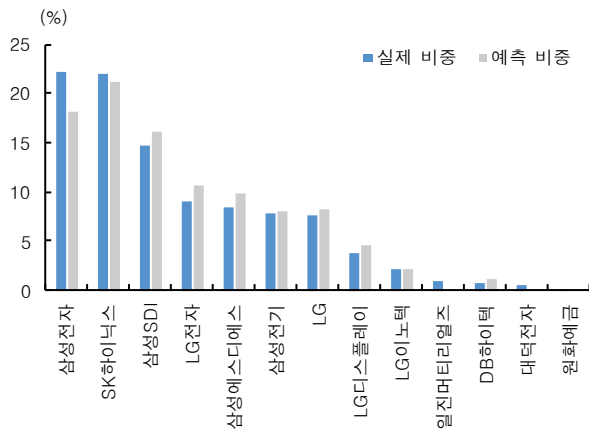
예측 종목과 목표 포트폴리오의 수익률 데이터를 가지고 선형회귀식을 구성할 경우, 회귀계수가 높은 종목일수록 목표 포트폴리오에 대한 영향이 커지게 된다. 이는 곧 목표 포트폴리오 내에 개별 종목의 비중이 높다고 해석할 수 있다. 따라서 회귀계수는 복제 포트폴리오에서 개별 종목의 중요도 또는 비중을 반영한다.



이러한 선형 복제방식에서 목표 포트폴리오와 복제 포트폴리오간의 트레이킹 에러를 최소화하는 최적화를 통해 종목 비중을 추정할 수 있다. 동시에 전체 종목 비중의 합이 1이 되도록 제약조건을 부가해 비중을 조정하게 된다.

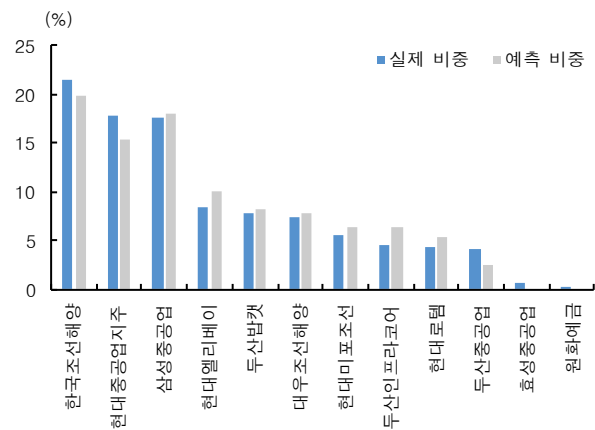
TIGER200 IT와 중공업 ETF의 경우에 실제 비중과 예측한 비중을 비교해 보았다. 실제 비중은 2019년 10월말 ETF의 종목 구성비이며, 예측 비중은 목표 포트폴리오의 기준가를 잘 추적하는 종목들을 먼저 선별한 후 회귀분석을 통해 추정한 비중이다. 실제 ETF 내에 비중이 높은 종목일수록 예측 비중도 유사하게 높아짐을 확인할 수 있다.

[그림 7] TIGER200 IT ETF의 실제와 예측 비중



자료: 미래에셋자산운용, 한국투자증권

[그림 8] TIGER200 중공업 ETF의 실제와 예측 비중



자료: 미래에셋자산운용, 한국투자증권

### III. 투자 시사점

#### 포트폴리오 복제를 통한 트레이딩 활용

포트폴리오 복제를 통해  
트레이딩 기회 확보 가능

알고리즘을 활용한 포트폴리오 복제가 투자에 갖는 시사점은 다음과 같다. 첫째, 목표 포트폴리오의 복제를 통해 트레이딩 기회를 확보하는 것이 가능하다. 목표 포트폴리오는 지수나 펀드 등 가격(기준가)이 공시되는 투자대상이면 가능하다. 펀드 기준가는 매일 공시되지만 보유종목에 대한 공시가 지연된다거나 일부 종목만 공개될 경우 유용할 수 있다.

펀드가 룡숏전략을 쓰거나 잦은 리밸런싱을 거치는 경우, 또는 매일 기준가 확인이 힘든 펀드의 경우에는 펀드의 성과를 트래킹하기 어려울 수 있다. 그러나 알고리즘을 활용한 포트폴리오 추론은 제한된 정보 하에서 효과적으로 포트폴리오를 복제할 수 있는 방법으로 트레이딩 기회를 확보하는데 도움을 줄 수 있다고 판단한다.

예를 들어 간혹 성과가 뛰어난 펀드의 경우에 해당 펀드가 보유종목을 공시하면서 일반 투자자들이 해당 종목에 대한 거래를 늘리고 가격도 영향을 받을 수 있다. 공시 이전에 펀드의 보유종목을 효과적으로 추정할 수 있다면 사전적인 포지션 구축을 통해 수익을 확보하는 것이 가능하다. 해외 헤지펀드를 복제하는 전략에 대한 연구가 많은 이유이기도 하다.

또한 펀드 매니저의 운용  
변화를 확인하거나 유동성  
높은 자산으로 복제하는데  
활용

둘째, 포트폴리오 복제를 통해 해당 펀드 매니저의 예상과 운용 행태 등을 간접적으로 확인할 수 있다. 예를 들어 지수 변경 이전에 포트폴리오 매니저가 어떤 종목의 편입을 예상하고 있는지 살펴볼 수 있다. 포트폴리오를 복제하기 위해 관찰 윈도우 기간을 이동해 나간다면 포트폴리오 매니저의 종목 편입비 변화를 확인할 수 있다. 이 때 매니저가 지수의 종목 변경에 앞서 선제적으로 실제 편입가능성이 높은 특정 종목의 편입비를 늘린다면 복제 포트폴리오에서도 해당 종목의 비중이 높아질 것이다. 또는 결산 시점 이전에 펀드 매니저가 포트폴리오를 조정하는 윈도우 드레싱 여부를 확인하는 것도 가능하다.

셋째, 펀드가 유동성이 낮은 자산으로 운용이 되고 있다면 동일한 방법을 거쳐 유동성이 높은 자산으로 복제하는 것이 가능하다. 유동성 프리미엄과 같은 초과적인 수익이 낮아질 수 있지만 복제 포트폴리오를 통해 경제성, 환금성, 신속성 등을 확보할 수 있다.

- 본 자료는 고객의 증권투자를 돕기 위하여 작성된 당사의 저작물로서 모든 저작권은 당사에게 있으며, 당사의 동의 없이 어떤 형태로든 복제, 배포, 전송, 변형할 수 없습니다.
- 본 자료는 당사 리서치센터에서 수집한 자료 및 정보를 기초로 작성된 것이나 당사가 그 자료 및 정보의 정확성이나 완전성을 보장할 수는 없으므로 당사는 본 자료로써 고객의 투자 결과에 대한 어떠한 보장도 행하는 것이 아닙니다. 최종적 투자 결정은 고객의 판단에 기초한 것이며 본 자료는 투자 결과와 관련한 법적 분쟁에서 증거로 사용될 수 없습니다.
- 본 자료에 제시된 종목들은 리서치센터에서 수집한 자료 및 정보 또는 계량화된 모델을 기초로 작성된 것이나, 당사의 공식적인 의견과는 다를 수 있습니다.
- 이 자료에 게재된 내용들은 작성자의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다.