Index Tracking in the Structure of Fund of Funds based on Cointegration

# Abstract

Index tracking funds have grown significantly in previous decade and attracted more and more investors as an outperforming passive investment vehicle. There are two main different ways to track indices. One is called full replication, funds can take long position on all the constituents as the same weights of an index. The other tracking method is known as sample replication, funds only buy part of the stocks from a family of index stocks using different analytics tools like correlation, mixed integer programming and cointegration. In this paper, our goal is to construct a portfolio to track S&P 500 in a structure of fund of funds (FoF) using cointegration analysis. In contrast with traditional index funds, we do not buy constituent stocks directly to mimic index, we buy sector ETFs. S&P 500 consists of 11 different sectors and industries, there are numerous sector ETFs on the market. We can construct an index fund by purchasing sector funds to track S&P 500 deploying cointegration analysis to make sure long run equilibrium. In FoF structure, we can cut transaction cost enormously and reduce turn over rate, which are essential for an index fund. Index funds hold stocks directly for both full and sample replication methodologies.

Contents

[Abstract 1](#_Toc518339352)

[Introduction 3](#_Toc518339353)

[Literature Review 6](#_Toc518339354)

[Data 9](#_Toc518339355)

[Methodology 11](#_Toc518339356)

[Research Design 16](#_Toc518339357)

[Portfolio Assessment 19](#_Toc518339358)

[Empirical Results 19](#_Toc518339359)

[Conclusion 19](#_Toc518339360)

[Limitations and Extensions 19](#_Toc518339361)

[References 19](#_Toc518339362)

# Introduction

In investment industry there are two opposite equity portfolio management philosophies, active portfolio management and passive management. Active investment is aiming to beat the benchmark index based on professional analytics and fund manager’s judgment in picking securities and determining the right moment to long and short them. Hedge fund is a typical active management portfolio. However, it is not as easy as planed in theory for active funds to outperform respective benchmarks. Based on SPIVA® U.S. Scorecard, around 83.18% of all domestic funds underperformed their benchmark for the 10-year period by middle year of 2015 in the US. Besides us stock market, many active funds failed to beat the targeted indices over the 10-year period in other major capital markets. On the contrary, passive investment pursues the same performance as targeted index over a long period of time instead of trying to beat the benchmark. A tracking fund is a classic passive product whose mission is to mimic a specified benchmark passively with buy-and-hold strategy, the benchmark can be a stock index, a commodity, bonds, even bitcoin. As opposed to underperformed active funds, index tracking funds attracted more and more investors and grown significantly through nearly a decade bull market since the worst situation in early 2009. Plenty of capital flow into passive funds rapidly. For 2017, investors poured more than $692 billion into index funds across all asset classes. For the same period, actively managed funds experienced $7 billion in outflows. Now the total asset in index funds including index mutual funds and index ETFs is about 1112 trillion in the US.

There are two main conventional methods to track indices.

One is called full replication, the fund can take long position on all the constituents of an index in the respective weights with buy-and-hold strategy, which is straightforward to implement and can achieve the precise tracking performance as long as fund managers rebalance the weights once a while. Even though full replication can closely track indices in theory, it has a few nonnegligible flaws. Full replication funds need to rebalance quite often with high volatility stock weights, which could lead to inflated costs. Liquidity is another issue, especially for small capitalization stocks, this may affect fund construction and increase the transaction costs. Low cost is a signature characteristic of passive management funds, but full replications funds cannot bring out this feature.

The other traditional tracking method is known as sample replication. Some indexes may contain large number of constituents, such as S&P Global 1200, Russell 2000. In those cases, full replication approach is not efficient to conduct, however, sample replication methodology can be appropriate. Sample replication funds need to long part of total stocks that could represent the underlying index based on correlations, risks and returns. As the funds trade a relative fewer constituents, which could significantly reduce the costs, but this may potentially cause higher tracking errors. <此处应有引用谁是最早提出了各种tracking 方法>

In additional to traditional physical funds holding a portfolio of assets, there is another alternative approach so called synthetic portfolio to replicate the performance of an index by using corresponding derivative and swaps instead of holding stocks directly. Proponents claim that synthetic funds are a better financial instrument than traditional tracking funds to track illiquid indexes at a low cost and small tracking error. However, the synthetic portfolios are born with a few risks involving counterparty risk, liquidity risk, and collateral risk. Synthetic funds are not popular in US markets due to regulation by US Securities and Exchange Commission.

In this paper, our goal is to construct a portfolio to track S&P 500 Total Return Index (SPTR) on the fund of funds(FoF) structure by using cointegration analysis that is a powerful econometrics tool that could ensure the long run equilibrium relationship between portfolio and the underlying index. As an approach of sample replication, we buy sector and industry ETFs to mimic SPTR index which consists of 11 different sectors and industries. There are numerous sector ETFs on the market, many of them have over a decade history, large asset size and high liquidity. We are going to select about 5 adequate ETFs from each sector to form a ETFs pool, then we design our portfolio based on LASSO regression to select ETFs and find corresponding weights in the co-integration system.

This paper is organized as follows. Section 2 reviews other literatures and compares different tracking methodologies. Section 3 describes targeted index S&P 500 Total Return and sector ETFs on the market. Section 4 introduces our tracking methodology cointegration and variable selection asset allocation method LASSO regression. Section 5 shows strategy implementation and empirical results for our tracking portfolios with varying number of ETFs and different rebalance strategies. Finally, section 6 makes the conclusion for our tracking strategy. In the end, we will discuss the limitations and potential research extensions based on this paper.

# Literature Review

Alexander (2005) is a pioneer who applied cointegration to passive portfolio management field. This paper deployed cointegration analysis to track a stock index, then built the long short market neutral strategy based on index tracking. The purpose of using cointegration is to identify any common stochastic trends in stock prices, and then achieve stationary tracking errors between a portfolio of stocks and the stock index over the long run. The author divided the process of constructing index tracking portfolio into two parts, selection and allocation. This paper took ‘brute force’ approach to select stocks. Firstly, pick the number of stocks to form the portfolio, then use all the combinations of stocks as possible portfolios. Next step is to optimize the weights of each stocks from every possible combination by using Engle-Granger cointegration methodology. This paper amplified the ordinary index tracking to long short market neutral strategy, which consists of a long portfolio tracking index plus, and a short portfolio tracking index minus. This long short strategy, as one of statistical arbitrage strategies, could provide double alpha opportunities in stock markets. Vast back testing results confirmed that Engle-Granger cointegration is a sound methodology to build index tracking portfolios with relative few stocks and less turnover rates.

Glova, Pastor and Sabol(2015) studied cointegration as a time series model and discovered its application in passive portfolio management. They discussed the statistical characteristics of cointegration and compared it with correlation from asset management perspective. They noted that cointegration and correlation are related, both describe the relationship between assets. Cointegration is a long-term relationship among time series. If cointegration existing, then it could ensure long run equilibrium between stock prices. Correlation is a short time statistic based on assets’ returns, that is not appropriate for constructing a long term buy and hold strategy. This literature tracked Dow Jones Industrial Average Index and Dow Jones Composite Average Index by exploiting the mean reverting property of cointegration. They used daily closed prices of indices and daily closed prices of component stocks adjusted for splits and dividends from 2000 to 2013. This paper conducted a lot of portfolios from different selection process and compared each return and risk metrics. At the end, they approved that cointegration is a right apply in passive portfolio management, which can create a comparable low volatility and low-cost tracking portfolio.

Sant’Anna, Filomena and Caldeira (2017) compared cointegration and correlation methodologies in index tracking and enhanced index tracking on Brazil Ibovespa index and U.S. S&P 100 index. This paper pointed out that both methodologies are outperforming for index tracking portfolios, but no significant advantages turn towards neither method for enhanced index tracking. The authors constructed a series of portfolios consists of at most 10 stocks by different combinations between in sample and out of sample data intervals through both approaches. Then they found different patterns between Brazilian and U.S. stock markets. There is a trade-off between tracking performance and costs in Brazilian market, which is correlation based portfolios have larger average tracking errors, but smaller turnover values, on the other hand, cointegration based portfolios have smaller tracking errors, but higher turnover rates accompanied higher cost. However, no empirical evidences revealed the similar features on S&P 100 index, tracking results did no favor neither correlation nor cointegration.

Overall, this paper failed to find robust evidences to demonstrate different characterises of cointegration and correlation in passive portfolio management area. It is worth noting that all portfolios have only 10 assets, which may be a potential reason why this paper did not generate strong findings. We will build portfolios with relatively more stocks and compare between numbers of stocks.

Numerous studies proved that cointegration is a sound and robust methodology to track index. However, there are another one crucial problems affect the tracking performance, construction costs, rebalance costs: asset selection. Asset selection is a picking art for fund managers. For our index tracking funds, assets selection helps to selection appropriate subset of assets out of total assets pool to represent the index, moreover, we can allocate different proportion of total capital to each asset in the portfolio. Alexander (2001)(cointegration and asset allocation) proposed ‘brute force’ approach to select assets. The author tested all possible combinations of a fixed number of stocks in a portfolio.

Therefore, brute force method requires huge computing power. When we track some other indexes contain large number of constitutes, it may cause explosive growth of computing, so that it is not applicable. For instance, the Russell 2000 index has 2000 stocks, if we want to pick 1200 stocks to construct a tracking fund, we may have N!/k!(N-k)!, the number of combinations is incredible giant, then the brute force becomes computationally infeasible even using modern supercomputers.

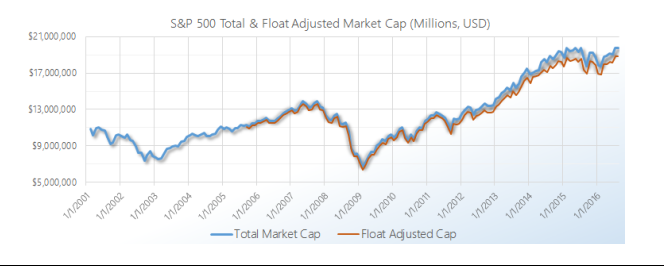
There are countless number of approaches to the asset selection problem. For linear regression, penalty methods are widely used as an effective statistical modelling technique. Regression with L1 penalty term is known as least absolute shrinkage and selection operator (Tibshirani, 1996), LASSO for short. For more details, we will discuss it in methodology section. Many academyes tested the tracking funds by using nonnegative-lasso method such as Wu et al. (2014) and Wu and Yang (2014). Yuehang Yang, Lan Wu (2016) proposed a two-stage nonnegative adaptive lasso method to do asset selection in ultra-high dimensional regression models based on adaptive lasso algorithm proposed by Zou (2006), which can deal with hundreds even thousands of stocks. They tracked CSI 300 Index that is a major index in Chinese stock market by using long and hold sample replication strategy. They did not use cointegration to ensure the long run equilibrium between tracking fund and CSI 300 Index. First stage solved the asset selection problem, they used nonnegative adaptive lasso method to select 30 stocks out of total 300 stocks, the number 30 is a predetermined number. Once they determined the stocks, second stage solved the asset allocation problem. They applied nonnegative OLS method to estimate the weights of the 30 stocks in the tracking fund. The authors did not show long time tracking performance, the results for short time were satisfactory.

# Data

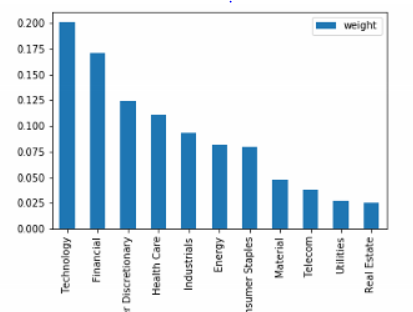
In this paper we constructed a fund to track S&P 500 Total Return Index(SPTR) by using sector ETFs. We used daily close prices of SPTR and these ETFs adjusted for paying dividends and stock splits, and the adjusted closing prices are easy for us to perform analysis of historical returns. We took 10-year data from beginning of 2008 to end of 2017, this period contains 2518 trading days, downloaded data from Yahoo! Finance.

S&P 500 Total Return Index is a very similar index to standard S&P 500 index. They have the same components that comprises 500 large capitalization companies listed on NYSE and NASDAQ and these constituents are categorized into 11 sectors: consumer discretionary, consumer staples, utilities, technology, health care, financial, energy, telecom, industrials, material, and real estate. Both SPTR and SPX are calculated in capitalization-weighted method in which the constituents are weighted based on the market value of their outstanding shares, however, the only difference is SPTR tracks the stocks with dividends are reinvested back into underlying stocks instead of tracking stock price movements only.

The total market capitalization of SPTR was about 23 trillion dollars at the end of 2017, which could cover over 80 percentage of US stock markets. Below figure 1.1 is the historical total market capitalization of constituents in SPTR, it could be considered as a successful representation of US stock markets and a critical indicator to reflect business cycle.

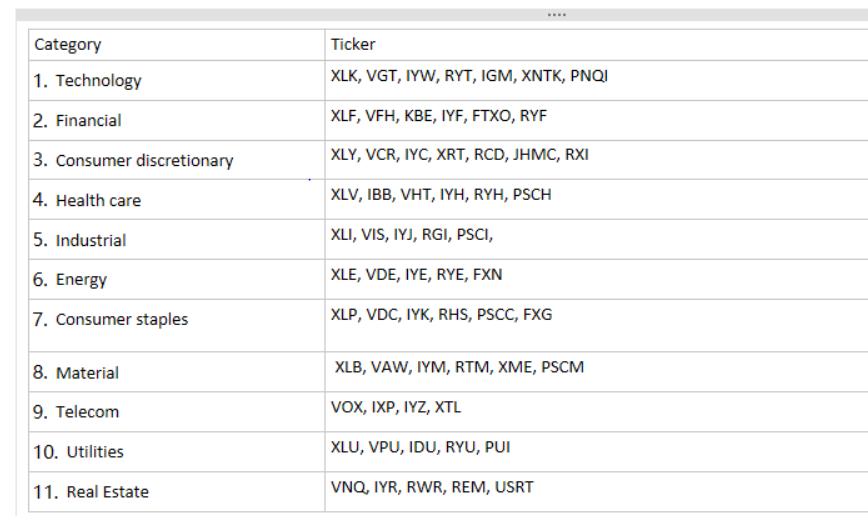


The 500 constituent companies are divided into 11 sectors, and the weights of different sectors change over the time. Figure 1.2 is a snapshot of sector’s relative percentage of total market capitalization of SPTR. The technology sector is the biggest, it has about 9 trillion dollars market capitalization, it weights up to 20%. The second position is financial sector, the proportion is about 17%, and the financial sector used to be largest for a long time. As we can see, the utility and real estate are the two smallest sectors, both sectors have about same capitalization, only contribute about 5% of total market capitalization. It is worth noting that real estate was emancipated from financial sector as a standalone sector since September 2016.



Benefit from the development of passive investment industry, there are plenty of EFTs focus on single sector. For simplicity, we chose about 5 ETFs for each sector to form our ETFs pool. We chose ETFs with outstanding tracking performance, good liquidity, relatively large asset size, and long history.

Below, table 1.3 is our ETFs pool.



# Methodology

Cointegration is a widely used time series methodology to identify and utilize the property of many time series that share a common stochastic trend and the long run equilibrium relationship among them. The concept of cointegration was initially suggested by Granger (1981), then in a seminal paper, Engle and Granger (1987) developed cointegration estimation procedures and tests.

In time series econometrics, the most important concept is stationarity, we usually build models based on stationary time series processes. Unfortunately, original stationary time series are very rare in finance. We deploy weakly stationary time series in our paper. Weakly stationary process is a stochastic process whose mean and variance are finite and do not change over time, denoted by I(0). That is:

A series {Xt} is weakly stationary if

1. E(Xt) = mu, mu is independent of time t
2. Var(Xt) = sigma(X)^2, is finite constant and independent of t
3. Cov(Xt, Xt-s) = gamma(s) is independent of t for all s

The financial returns are weakly stationary time series process; nevertheless, many financial data like stock prices, interest rates, exchange rates are not stationary processes. A non-stationary time series {Xt} is called integrated of order d, if it can be made stationary by differencing d times, denoted by X ~ I(d).

Typically, log stock prices are a random walk with drift, integrated of order 1 I(1), we build our model by using log prices where the logarithm transformation could give us the continuous return and other good properties.

Xt = mu + Xt-1 + epsilon t epsilon t ~ iid(0, sigma^2) is the white noise, The constant mu is called the drift, Mu = E(Xt-Xt-1), which is the time trend of log prices.

As for multivariate time series, each of them is integrated of the same order, but some linear combination of them has lower integration order, then we can say these time series are cointegrated.

More general form is for a N dimensional variable X:

if 1. Xi ~ I(d), i = 1,2,..N

2. exist 1 or more linear combinations Zt=alpha’ Xt s.t. Zt ~I(d-b) b >0

Then X ~CI(d,b)

For example, both Xt, Yt are random walker processes I(1), Zt=Xt-alpha\*Yt ~I(0). So Xt Yt are cointegrated, and the coefficients [1, alpha] is called cointegration vector. Cointegrated time series have may nice properties and characteristics, here Xt, Yt have the common stochastic trend, they will move together in the long run equilibrium state. And the Zt is the short-term deviation from the equilibrium, which is the mean reverting property we can benefit from.

When we apply non-stationary time series into a regression model, it may cause misleading statistical inferences among them, which is called spurious regression discussed by Granger and Newbold (1974). The regression model mistakenly provides a non-existing relationship between independent regressors and response variable with statistically significant coefficients and high R^2. Because two non-stationary stochastic processes move together does not mean they are related, it rises the spurious correlation.

Cointegrated time series can avoid the problem of spurious correlation. Since the cointegrated time series share the common stochastic trend and the linear combination is a stationary process, they will not deviate far away.

There are many methods to test the present of cointegration relationship in multiple time series, Engle and Granger (1987), Johansen (1988) and Johnsen and Juselius (1990). In this paper, we applied Engle Granger test method. For more information about estimates and tests of cointegration system in not part of this paper, we not talk about it in details.

The Engle Granger method tests the null hypothesis of no cointegration among the time series. Engle Granger method requires all the time series have the same number of integration, then essentially perform unit root test on the residual term of the cointegration regression. If residual term is stationary process, then the corresponding time series are cointegrated, and vice versa.

For instance, X and Y are I(1), we regress X onto Y by using ordinary least square.

Yt = beta0 + beta1 Xt + Et.

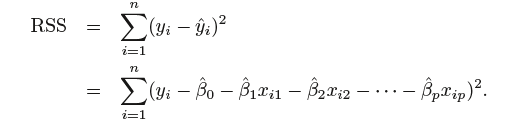
Then apply augmented Dickey-Fuller root test on Et. Null hypothesis of ADF test is the presence of unit root, which indicates that the time series is not stationary. Only residuals are stationary, we can conclude that X and Y are cointegrated.

Beside cointegration analysis ensures long run equilibrium state, we employ least absolute shrinkage and selection operator (LASSO) regression as a key technique to select proper assets and allocate our capital.

For a usual multiple linear regression:

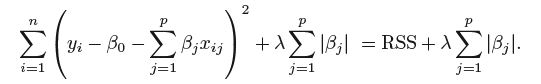
Screen Clipping

The classic methodology to estimate the unknown coefficients is ordinary least square(OLS) method, which is to minimize the sum of squares of the differences between response variable and each predicted value by the estimated function. The explicit form is:



OLS is a classic and mature algorithm to solve multiple linear regression, however, for high dimensional data, we want to get a model that is easier to interpret and less complexity by shrinking some coefficient estimates to 0, which is equivalent to exclude those variables.

There are many alternative approaches to do variable selection and shrinkage: best subset selection, stepwise selection, ridge regression, lasso regression and elastic net. In this paper, we adopted LASSO regression.

βL = where alpha >= 0.

Notice that for any alpha >0, there exists an s, equal to |beta apha hat| where:

Screen Clipping

This is the dual form of the optimization problem.

the only difference equation 2.1 and equation 2.0 is that the objective function contains a penalty term that is defined as the sum of the absolute value of coefficient estimates.

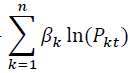
The  is called L1-norm, and the alpha is called tuning parameter that can control the shrinkage degree to coefficient estimates. When alpha is 0, the penalty term does not affect the objective functions, which is the same as normal OLS. On the contrary, as alpha is getting larger, the impact of the shrinkage penalty grows, the coefficients estimates are approaching 0. With penalizing the coefficient estimates, some variables are removed out of our model, so we get a subset of variables. As alpha = infinite, all the coefficients will be 0. Therefore, the choosing an appropriate alpha is critical to the model.

In general, there is no simple closed-form solution to lasso regression. There are several optimization methods to solve lasso regression, such as coordinate descent, least angle regression. In this paper, we applied coordinate descent algorithm by using Sci-kit Learn package in Python.

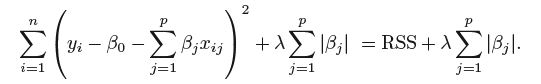
# Research Design

We introduced the cointegration and lasso regression as the powerful techniques that help us to build a fund of ETFs to track SPTR.

In the beginning, we need to determine the selections of ETFs in our fund and then optimize our capital weights for each asset. We employed a lasso linear regression of log adjusted close price from beginning of 2008 to end of 2014: the dependent variable is SPRT index, and the explanatory variables are the log prices of each asset from our ETFs pool. The model has the form:

Ln(SPTR\_t) = beta0+ + epsilon\_t

The coefficient estimates are estimated by the method of LASSO, it has the equation:

βL =arg min where alpha >= 0.

Subject to beta L >= 0.

Alpha is the tuning parameter in lasso L1 regularization, it could determine the values for coefficients estimates as we motioned in previous sector. It is important to select an appropriate value for tuning parameter in lasso, however, the optimal tuning parameter is difficult to calibrate in practice (Lederer and Muller, 2015). Fang and Tang (2013) note that “To the best of our knowledge, there is no existing work accommodating tuning parameter selection for general penalized likelihood methods.”

We treated the selection of tuning parameter as an endogenous problem. Starting as alpha = 0, then set a relative small step length tau=0.0001, every time add it to previous alpha to get a new alpha, until alpha equal to 1.

Alpha i = alpha i-1 + tau where i >=1, alpha 0 = 0, and tau = 0.0001

For every alpha, it will have a new lasso regression model, we will generate a different set of coefficient estimates, and underlying residual series.

Then we test cointegration relationship between SPTR index and the combination of non-zero ETFs by applying Engle-Granger cointegration method. We use Augmented Dickey Fuller method to test the presence of unit root in residual series. Screen Clipping

If the residual series have not unit root, which means it is stationary and we get long run equilibrium state from Engle Granger methodology. In contrast, if the residual series has unit root existing, we do not go along with the corresponding combination of ETFs.

Once we find the stationary residuals, indicates that we have selected the components of our fund.

Different alphas may lead to the same combination of ETFs but with slightly varying coefficient estimates, we chose the most cointegrated one as our fund. P-value from the ADF test on residual series is an indicator to reflect the stationarity, so it could be a good indicator to measure the degree of cointegration relationship between SPTR index and the assets combination. The smaller p-value, the stronger cointegration relationship.

Next issue is to allocate capital to each ETF in our fund. Because we adopted long and hold tracking strategy, we cannot short sell ETFs, the coefficients cannot be negative. From equation 3.2, the constraints require that all the coefficients estimates must be non-negative, zero means we exclude the ETFs.

As a common practice (Alexander 2008)to calculate the weights, we divided the coefficient estimates by the sum of total value of coefficients except the constant term beta0. The ratio of each components is the weight of it in our fund. In math form:

weight\_i = beta\_i / Screen Clipping and Screen Clipping wi >=0.

With stronger penalizing, more ETFs will be removed out of the combination. The regression model could provide cointegrated portfolios consists of relatively small number of constituent ETFs. We will construct and compare three funds with 15, 10 and 5 ETFs.

Cointegration methodology offers a rationale for tracking targeted index over the long term. Based on the theory, the cointegrated portfolio will tie together with the index in the long run, it may only deviate away from the index temporarily. Therefore, we do not need to rebalance the weights of ETFs in portfolio by the intrinsic characteristic of cointegration.

We will test the tracking performance of portfolios between no balance and different rebalance time intervals: annually rebalance, semi-annually rebalance and quarterly rebalance. For a cointegrated combination of ETFs with the initial weights, we will balance the weights of each asset by running regression equation, and then we also run ADF test on new residual series to make sure they are cointegrated. We have test period from beginning of 2014 to end of 2017. We will rebalance weights of constituents 3 times for annually rebalance, 7 times for semi-annually rebalance, and 13 times for quarterly rebalance.

Generally speaking, more frequently rebalance our portfolios could improve the tracking performance because we can adjust the funds short-term deviation from index manually. At the same time, rebalancing will produce huge transaction costs. The enormous transaction costs could ruin the advantages of the tracking funds, and make the passive strategies less appealing to inverstors . For simplicity, we assume the transaction cost is 0.1% of the tarding amount for institutions.

To sum up , we will build 3 profolios based on number of component ETFs, and for each portfolios, we have 4 different rebalance schemes, a total of 12 funds.

## Portfolio Assessment

We fitted a regression model and estimated the weights for underlying ETFs in portfolio in train set from 2008-01-01 to 2013-12-31. Our goal is to mimic the performance of SPTR and achieve the same profitability and volatility. We expected to minimize tracking errors and get a highly cointegration relationship between portfolio and underlying index. Then for out-of-sample period from 2014-01-01 to 2017-12-31, we operated our portfolios in market and then we took the information ratio to test it as well. We adopted a few measures to assess the tracking funds’ fitness.

* ADF test on residual series – measures the stationarity of errors from regression model that is the proxy of cointegration relationship. Smaller P-value indicates more stationary residual terms and stronger cointegration relationship between log prices of portfolio and SPTR index.
* Tracking errors mean and volatility -- the tracking error is defined as the deviation of daily returns between portfolio and SPTR. We would like to have small mean and low volatility tracking errors, which implies a stable and robust tracking fund.
* Correlation coefficient of daily returns –- We construct a cointegration relationship between the log prices of the portfolio and SPTR index, and the daily return is the first difference from log prices. This is used to directly reflect the tightness between daily returns from the portfolio and SPTR index.
* Information ratio -- it is commonly used to measure the risk-adjusted returns of a portfolio above a benchmark. It has the formula:

Screen Clipping

where Rp is the daily return of portfolio and Ri is the daily return of SPTR index. Positive and higher IR means outstanding risk-adjusted returns, and it is attractive for investors. However, we do not seek higher IR particularly as operating an index tracking fund.

# Empirical Results

# Conclusion

# Limitations and Extensions

# References