Final Project Report

ETF Portfolio Construction and Optimization

Course Code	STAT GU4261
Course Title	Statistical Methods in Finance
Section Number (if any)	001
Instructor	Zhiliang Ying
Project Title	ETF Portfolio Construction and Optimization
Group Number	12

Member Name	UNI	Contribution
Cao, Youyang	yc3232	ETF Clustering by K-means
Rong, Bo	br2498	Future Trend Prediction by Time Series
Gopinath, Rudra	rg2986	Portfolio Comparison
Guo, Zhiyi	zg2249	K-means Loss Function & ETF Research
Ji, Jessica	xj2197	Portfolio Comparison
Li, Katie	ml3749	Project Introduction & Conclusion
Lin, Xiaotong	xl2506	Portfolio Construction by Markowitz Theory

Words: 2530

Submission Date: May 9, 2017

1. Introduction

In the financial market, every investor may expect different returns and different tolerance levels of risk. How to wisely choose securities and adopt a good portfolio strategy is always the main topic in investment. During the last decade, Exchange Traded Funds (ETFs) are getting increasingly popular and their inflows are reaching the all-time high. This is due to several benefits offered by ETFs: 1) diversified nonsystematic risk over one field or multiple markets; 2) performing as the same direction as mutual funds, but offering lower transaction costs and taxed rates; 3) can be traded more frequently than mutual funds, like stocks, during a trading day. We want to take the advantage of all these excellent features of ETFs to construct portfolios that can fulfill different investors' needs, using the quantitative approach with Modern Portfolio Theory to construct portfolios by only training ETFs.

2. Objectives and Project Plan

Based on a dataset of 921 ETFs' daily close price within last 580 trading days from Nasdaq.com, we calculate daily log returns and process from there. By randomly selecting the returns within periods of 120 rolling days as training data, we use K-means method to cluster the ETFs into groups, and pick out two ETFs from each group that have the highest Sharpe Ratio. We then implement the Markowitz Portfolio Theory to optimize or maximize expected return based on a given level of market risk. Based on three levels of risk, we find the optimal allocations on the efficient frontier, which gives us the best return on a given risk level. Then we use the return of our portfolios for next 120 days as training data and compare it with 5 actively-managed ETF portfolios to check whether our portfolios outperform others. After repeating this process for 200 times, we can evaluate the portfolio performance based on the probability that our portfolios perform better than others.

3. Methodologies and Analyses

1) ETFs Clustering by Loss function and K-means

Our goal is to design portfolios that can meet the requirements for risk-averse, risk-neutral, and risk-seeking investors, by differentiating on risk and return levels. Our differentiating strategy will be introduced below:

Step 1: Update each c_i (assign data to cluster): we assigned each data point to the nearest cluster by minimizing the Euclidean distance. $\mathrm{i} = 1, 2 \ldots, \mathrm{n}$

Step2: update each μ_k (centroid): after the new assignments of data points to each centroid, we calculated the empirical mean of those data within each cluster and assigned the value to the corresponding centroid.

Step 3: Repeat the above two steps for 20 times.

Function used to assign data points to centroids and update centroids are shown as below:

Update each c_i: $c_i = argmin_k ||x_i - \mu_k||^2$

Update each μ_k: $n_k = \sum_{i=1}^{n} I\{c_i = k\}$ and $\mu_k = \frac{1}{n_k} \sum_{i=1}^{n} x_i$

Loss function: $\mathrm{L} = \sum_{i=1}^{n} \sum_{k=1}^{k} I\{c_i = k\} ||x_i - \mu_k||^2$

This is the objective function that we want to minimize which aims to minimize the Euclidean distance of every single data point to its assigned centroid.

When trying to cluster ETFs to several groups with different risk and return level, we implemented K-means algorithm to achieve the goal. The two dimensions of features we used in clustering are daily return and daily risk (standard deviation) with 120 trading days as a period

(Figure.1), including all 1132 ETFs in the pool. We chose 120 trading days as a trading period considering the transaction fees and also because of the later use of the return comparisons between our portfolios and the benchmark. Using the coordinate descent algorithm, we updated the centroids and data assignments iteratively, and calculated the objective loss function after each iteration. After 20 iterations, the objective function that we want to minimize is approaching a stable value and the decrease is minimal and negligible. In general, the loss function will be smaller by choosing a larger K. Also, we want to keep a reasonable amount of groups that could help us differentiate ETFs. We then tested K from 2 to 5 and the drop of the loss function (Figure.2) From each cluster, we choose 2 ETFs with the highest Sharpe Ratio. We used Sharpe Ratio as criteria because it represents the average return earned in excess of the risk-free rate per unit of volatility or total risk. It is believed that the higher Sharpe Ratio gives better ETF. After that, we ran the algorithm for 200 times for sub trading periods to construct different portfolios of 10 EFTs.

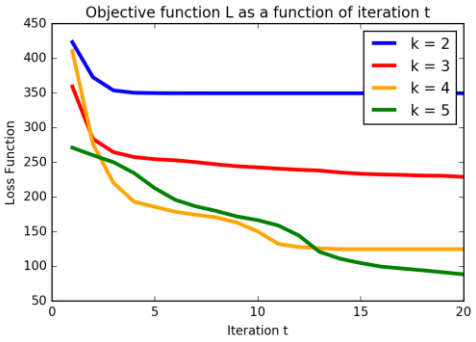

Figure.1: Loss Function

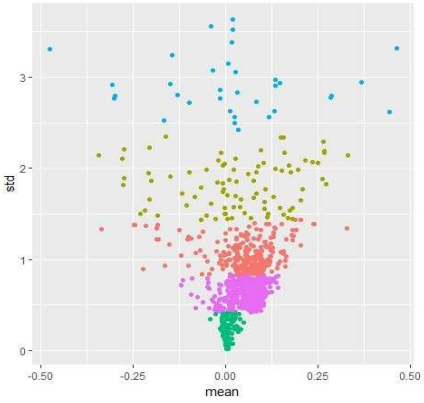

Figure.2: K-means Clustering

2) Portfolio Construction by Markowitz Portfolio Theory

After selecting 10 most ideal ETFs by K-means clustering, we want to figure out how much proportion of total fund should we distribute to each ETF to construct an optimal portfolio.

In 1952, Harry Markowitz developed the significant portfolio-selection technique, which became the foundation of modern portfolio theory (MPT) (Witt, 1979). Applying the theory, we are going to construct portfolios based on the following formulas.

Expected return for a portfolio: $\mathrm{E}(r_p) = \sum_{i=1}^{n} w_i \mathrm{E}(r_i)$, where $\sum_{i=1}^{n} w_i = 1$; n = the number of ETFs; w_i = the proportion of total fund invested in ETF_i; and r_i, r_p = the return on i[th] ETF and portfolio p.

Expected risk (variance) of a portfolio combination of ETFs:

$$\mathrm{Var}(r_p) = \sigma^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j Cov(r_i r_j)$$

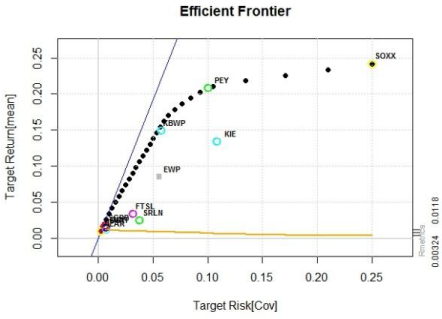

Figure.3

Since every possible ETF combination can be plotted in risk-return space, based on the data of latest 120 trading days, the collection of all such possible portfolios defines a region in this space. The line along the upper edge of this region is known as the efficient frontier, on which are portfolios providing the maximum return given specific levels of risk (Figure.3). As graph shows, the red point shows the minimum variance portfolio(MVP). This is the also the tangent portfolio, where the optimal curve touches the blue capital market line.

Next, we can calculate the weights of optimal risky portfolio by

$$\text{Min } \sigma^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j Cov(r_i r_j)$$

Subject to $\sum_{i=1}^{n} w_i E(R_i) = e$, where e is the target expected return, and $\sum_{i=1}^{n} w_i = 1$.

By utilizing the method of Lagrange multipliers, the weight for each ETF of the portfolio with targeted return can be determined. Later, we can also use R to draw weights for each ETFs of portfolios with different targeted return (Figure.4). In the graph, every vertical line represents an optimal risky portfolio and the dark black one shows the MVP. Briefly summarize, the MVP's return is about 0.95% annually, and its risk it about 0.26%. In risk-adjusted terms, this is rather impressive. However, noting that the MVP has heavily weight of iShares Short Maturity Bond (NEAR), which shows in orange. Since the optimization above is unconstrained, not surprisingly, the best choices tend to be extreme as the most optimal ETF take a huge proportion of the portfolio.

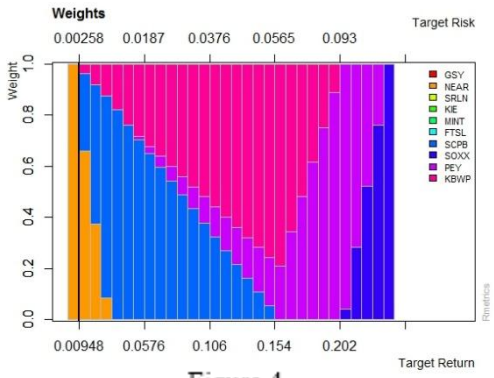

Figure.4

Therefore, suggest we want constrained portfolios with minimum weight 5% and maximum weight 50% of every ETF. By adding the constraints, portfolios seem to be moderated. The result shows that the MVP's return increases up to nearly 5% annually, but its risk also increases to about 3% (Figure.5). Significantly, weights of iShares Short Maturity Bond (NEAR), the orange ones, are much lower than previous ones.

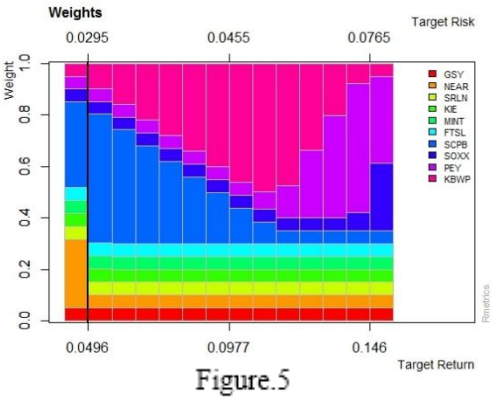

Figure.5

Besides constructing optimal portfolios, different investors might prefer portfolios based on different levels of risk, so we continue constructing three types of portfolios as shown in (Figure.6) below.

Portfolio Type\ETF Name	GSY	NEAR	SRLN	KIE	MINT	FTSL	SCPB	SOXX	PEY	KBWP	Target Return(%)	Target Risk(%)
Conservative Portfolio (%)	5	5	5	5	5	5	38	5	5	22	7.4	3.7
Moderative Portfolio (%)	5	5	5	5	5	5	20	5	5	40	9.8	4.6
Aggressive Portfolio (%)	5	5	5	5	5	5	5	5	13	47	12.2	5.5

Figure.6

Only based on the target return and risk above, it is hard to determine how good are our portfolios. Hence, we are going to test our whole algorithm for 200 times and compare the performance of our specific types of portfolios with other same types of excellent portfolios in the market. Each time we will use data of different 120 trading days to construct portfolios with 10 distinct EFTs.

4. Results and Interpretations

1) Research on Most Common ETFs

By running the K-means algorithm for 200 times, we got 200 portfolios consisting of 10 ETFs. We picked out 5 ETFs that are appearing most frequently among those portfolios and took a closer look at them to understand what makes them stand out. The five ETFs are: Guggenheim Enhanced Short Duration ETF (GSY), iShares Short Maturity Bond ETF (NEAR), SPDR® Blackstone / GSO Senior Loan ETF (SRLN), PIMCO Enhanced Short Maturity Active ETF (MINT), First Trust Senior Loan ETF (FTSL). We found that these 5 ETFs share some common investing strategies. For example, most of them seek maximum current income and preservation of capital and daily liquidity. In addition, they achieved diversified exposure to short-term bonds with an average duration of less than one year. The short duration means lower interest rate risk and sensitivity to changes from the Fed. Also, most of the ETFs have a majority of net assets allocated to cash, U.S. investment-grade

fixed-income securities, or senior loans, which enables higher liquidity and lower default risk with first in priority in receiving payments from a borrower.

2) Comparison with Benchmarks

To measure the performance of our portfolios, we decided to compare them against several actively-managed ETF portfolios in the market. The benchmarks we choose are the five largest actively managed mutual funds by Assets under Management (AUM), including PTTAX, VBMFX, VTSMX, AGTHX, and VFINX. We separated the five benchmark funds into three groups (conservative, moderate, aggressive) based on their assets compositions and compared with our three risk types of portfolios respectively.

As mentioned above, we divided all daily price data into 200 subsets with 120 trading days in each. We used each subset to construct 3 types of portfolios and compare the performance of our portfolios with 5 other benchmarks' by using data of the next following 120 trading days. Finally, we calculate the percentage that our portfolios beat the benchmarks among 200 trials as the performance metric shows. The Figure.7 below shows the specific results taken from sampling trials:

Conservative (total bond fund)		Moderate (stock market index fund)		Aggressive (growth equity fund)
PTTAX	VBMFX	VTSMX	VFINX	AGTHX
65%	59%	48%	70%	61%

Figure.7

The numbers show our portfolio strategy beat the benchmarks with similar risk, over half of the times. The comparison results indicate a good performance of our ETF portfolios and validate our strategy.

5. Conclusions and Discussions

1) Portfolio Price Prediction by Time Series (ARIMA Model)

In statistics and econometrics, and particularly in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. Both models are fitted to time series data either to better understand the data or predict future points in the series.

In the ARIMA model, the future value of a variable is a linear combination of past values and past errors, expressed as follows:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_{p-1} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

where, Y_t is the actual value, ε_t is the random error at t, φ_i and θ_i are the coefficients, p and q are integers that are often referred to as autoregressive and moving average, respectively.

ARIMA	AIC
(1,0,0)	197.1326
(2,0,0)	199.0837
(0,0,1)	365.6868
(0,0,2)	280.1512
(1,1,0)	195.1001
(0,1,0)	193.1028
(0,1,1)	195.1001
(1,1,2)	199.1001
(2,1,0)	197.0999
(2,1,2)	195.8336

Figure.8

To determine the best ARIMA model fit each ETF portfolio, the AIC criteria are used for each ETF. The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

Figure.8 is the result of the different ARIMA parameters for Conservative Portfolio, IMA (2, 1, 0) is considered the best model for Conservative Portfolio. The model returned the smallest Akaike information criterion of 193.1028.

We use the data of the most recent 120 days to predict the Conservative Portfolio price of the next 15 days by using ARIMA (0,1,0) model. As Figure.9 shows, the forecasts are plotted as a blue line, the 80% prediction interval as a dark shaded area, and the 95% prediction interval as a light color shaded area.

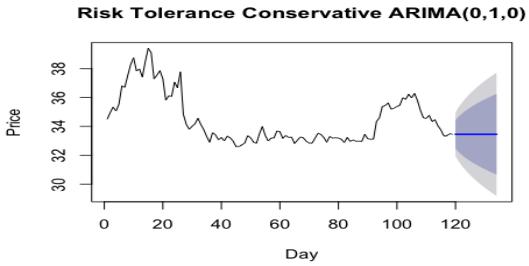

Figure.9

Similarly, the ARIMA (0, 1, 0) is considered the best model for Moderate Portfolio as well. The model returned the smallest Akaike information criterion of 42.51725. We use the data of the most recent 120 days to predict the Moderate Portfolio price of the next 15 days by using ARIMA (0,1,0) model. The forecasts are plotted by Figure.10.

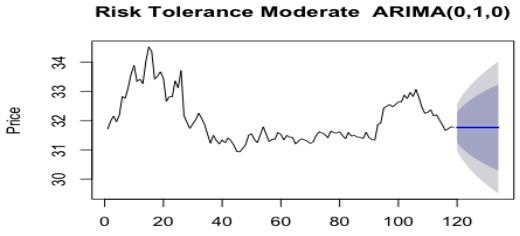

Figure.10

The ARIMA (2, 1, 2) is considered the best model for Aggressive Portfolio. The model returned the smallest Bayesian or Schwarz information criterion of -178.9682. We use the data of the most recent 120 days to predict the Aggressive Portfolio price of the next 15 days by using ARIMA (2,1,2) model. As Figure.11 shows, the forecasts are plotted as a blue line, the 80% prediction interval as a dark shaded area, and the 95% prediction interval as a light color shaded area.

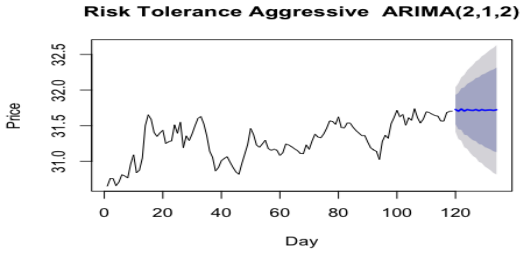

Figure.11

Generally, Time Series help us make a rough prediction about our portfolios' future price trend; however, we still need more data to make our forecast more accurate.

2) Conclusion

To control the transaction fee of ETFs, we need to avoid over-selecting ETFs. Because of that, we think 5 is a reasonable number of clusters which has a loss function attaining a reasonable low value after 20 iterations in the K-means analysis. By selecting two ETFs from each cluster based on their Sharpe Ratio, we end up with a total of 10 ETFs to construct portfolios. According to the Markowitz Portfolio Theory, we select portfolios in a 120-day period that have the best return based on three levels of risk, 3.7%, 4.6%, 5.5% respectively. In this way, we can divide the potential investors into conservative, moderate, and aggressive groups based on their risk bearing ability.

After changing our 120 trading-day dataset by every 3 days, we repeat constructing portfolios for 200 times, and we are able to compare their performance with the five most actively managed portfolios. It is impressive that our portfolios outperform the other popular ones over 60% of the time. Finally, we also use Time Series Analysis to predict the future trend of our portfolios' performance and give our investors a general idea of what will be a reasonable expectation.

Certainly, there are some limitations on our strategy. To further improve our project, we will consider following points in the future: 1) When conducting the test for portfolio comparison, we divided the data into 200 subsets with 120 days in each. Because we only have a total of 580 trading days in total, there are many overlapping between each subset, causing some variance in our results. Hopefully, we can use a much larger data pool to eliminate the repeating use of the same period data. 2) Instead of setting 120 rolling days as the training period, we can use cross-validation to select the best size of the training data that can give us the best test result. 3) We would like to have a control of the updating period of our portfolios. All the insights above can be taken into consideration for out project's future refinement to achieve a better performance and cater to more potential investor's preference and goal.