Research on the Investment Value of Stock Exchange ETF in China Based on Factor Analysis and Cluster Analysis

Zhaobei Wei*

Department of Mathematics, Stevens Institute of Technology, Hoboken, USA

*Corresponding author: zwei21@stevens.edu

Abstract. A well-managed portfolio with thoughtful investment selection depends on the market. It is challenging to manually track share market prices because they fluctuate based on a variety of reasons. This article sorts out the traditional theory of domestic research fund performance evaluation, and selects 12 indicators suitable for analyzing ETF performance in China for factor analysis. The selected 629 sample funds are listed stock ETF issued from 2015 to 2020 and extended to 2022. Clustering methods used is k-means clustering with a Euclid distance and with shape-based distance measure. Factor analysis shows that the current ETF fund research indicators can be divided into 4 categories, namely income factor, anti-risk ability factor, management ability factor and volume factor. According to these three types of indicators, samples are grouped into 7 categories, and the study shows that most of the listed stock ETF funds have good risk diversification ability, and the performance of various funds varies greatly.

Keywords: ETF; factor Analysis; cluster Analysis; investment value; fund performance.

1. Introduction

ETFs are the main investors in the financial markets for the resource allocation. It has far-reaching significance for the transformation of investment philosophy, affecting the change of market structure and investment conduct [1,2]. Considering relative performance differences for funds of different sizes, Daniel, Christine, Micheal and Wayne explored the relationship between fund size and fund performance [3]. The empirical results show that when the size of funds managed by active funds reaches a certain level, their costs can be completely covered by income. But when the fund size exceeds the optimum size, its benefits decrease as scale increases. Kothari and Warner regard fund assets, standard deviation and β index as input indicators, and fund returns as output [4]. The result shows that funds with low beta value and small size operate more efficiently. Olson and Dellva analyzed that if the fund is in front-end charging, it will perform worse in terms of risk-adjusted return [5]. Dowen and Mann concluded that in order to generate returns, under-performing fund managers will trade more, but at the same time they will incur more transaction costs, thus leading to worse performance [6].

Since 2004, the ETF fund market has become a favorable addition to financial of China. In the field of open-end fund research, Chinese scholars offer different things Research ideas and perspectives. Chen Yongsheng and Yang Ning used three traditional evaluation indicators such as Sharpe index and Treynor index score to analyzing the performance of Chinese funds, the research concludes that the performance of Chinese funds is still better than the market portfolio after risk adjustment, but these three traditional evaluation methods do not have significant correlation [7]. It can be seen that the banking and insurance industries have good investment value, while the investment value of securities and other financial industries is relatively low, and note that the advantages and disadvantages of different companies are also different. Listed companies in the banking industry have good solvency, profitability and growth and operational capability [8]. Also, IT companies bring a delightful market to investors. Taking Tencent Holdings as an example, the company's financial indicators are from five aspects: the company's growth, cash flow, profitability, operation ability and solvency ratio [9]. While Zhang Tiancheng's research on the stock field shows that the indicators seen in the research report of sellers in security companies appear to be effective only in reports [10].

But in general, the study of ETFs is still in its infancy and needs to become a more systematic evaluation system. Accordingly, this study is based on a factor analysis model, demonstrate the current state of the ETF fund market and provide regulatory ETF markets ideals for investors. The idea of the field will help form a more scientific and healthy ETF market.

2. Data and Algorithms

2.1. Normalized

First, since we different types of data, variant dimensions may influence the analysis. Therefore, data sets need to be normalized to eliminate indicator differences between units and other aspects effect. Hence, we have the following:

$$x_{m,n} = \frac{x_n - \min(x)}{\max(x) - \min(x)}. \tag{1}$$

Where m is the indicator, n represents the nth number of the indicator.

2.2. Principal Component Analysis

A statistical method called the principal component analysis is used to examine how many variables interact with one another and to explain those interactions in terms of a smaller set of variables (i.e., principal components). Using the correlation matrix for the stock watch portfolio, eigenvalues and eigenvectors are calculated, which implies corresponding principal components. Principal component coefficients display eigenvectors for the principal components. Components with an eigenvalue of less than 1 are omitted. Values having a strong connection between the principal components and the original standardized variables are essential in regard to the scenario threshold.. The number of components matches the number of unique stocks in a stock portfolio. Because PCA uses the correlation matrix, the variables are standardized: Values along the correlation diagonal are equal to 1, and the total variance is equal to the number of unique stocks. Var variable and Cumulative variable display percent of variance for each principal component and percent of variance for each principal component. The Scree Plot displays the eigenvalue for each principal component by the number. The latter components cause the line to become almost flat, showing that their contribution is increasingly negligible.

2.3. K-means Clustering Method

In order to identify a natural grouping of data, if any, cluster analysis uses the information provided by the analyst in the form of pertinent qualities. It is crucial to remember that using cluster analysis as a data-mining method is pointless because expert subject knowledge is a crucial component of effective clustering. Discovering structure and relationships in data is made possible through the use of cluster analysis. A cluster analysis' findings can directly influence the creation of classification schemes. In actuality, a set of results only relates to the sample on which they are based; however, by appropriately modifying the technique used, it can be expanded to correctly represent the qualities of additional samples and ultimately the parent population.

Contrary to discrimination analysis, cluster analysis does not rely on prior knowledge of significant distinctions within a population. With the exception of cluster analysis, every data analysis technique has its roots in a certain field. Psychology is the field that invented factor analysis and other scaling techniques. Although regression is utilized in many different fields, econometricians have produced a sizable body of literature on the method. Numerous disciplines have separately used cluster analysis (engineering, econometrics, psychology, and biology). Punj and Stewart provide a thorough review of cluster analysis's uses in social science [11]. The next paragraphs outline the cluster analysis's steps.

The method, created by MacQueen in 1967, uses a Euclidean distance measure to divide a sample of n entities into k sets. Each item is assigned by the algorithm to the cluster with the closest mean.

The following three steps can be used to explain the process. First, divide the objects into k initial clusters. Afterward, go through the list of objects in the data set, assigning each one to the cluster whose mean is closest (one of the most common distance measures is the Euclidean measure, which may be used with either standardized or unstandardized observations to calculate the distance). The mean should be recalculated for both the cluster that gains the new item and the cluster that loses it. Finally, keep repeating Step 2 up until there are no more reassignments.

One could define k initial centers and then move on to Step 2 instead of starting with a division of all items into k preliminary groups in Step 1. Each observation in the data set is treated by K-means as an object with a location in space. The partitioning of the objects places them as distant from objects in other clusters as possible while keeping them as near to one another as possible inside each cluster.

The member objects and the center of each cluster in the partition serve as defining characteristics. The distance from all of the objects in a cluster added together is minimized at the center of each cluster. A few restrictions apply to the K-means approach. When choosing the initial number of clusters, attention must be taken to avoid local minima and misclassification.

3. Results and Discussion

As can be seen from Table 1, the sig value of Bartlett's Test of Sphericity is less than 0.05. Then, at a significance level of equals to 0.05, the null hypothesis is rejected and the correlation coefficient matrix is considered has significant differences with the identity matrix. Also, with Kaiser-Meyer-Olkin Measure of Sampling Adequacy between 0.6 and 0.7, the data set is counted to be available using principal factor analysis. Hence, the original variables are judged to be suitable to do factor analysis according to the metric performance.

Table 1. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.681
Bartlett's Test of Sphericity	Approx. Chi-Square	5288.107
	df	45
	Sig.	.000

With the help of the factor analysis method, we could verify the relationship among the variables and the whole data set and decide which is more significant. By plot the PCA screes in Figure 1, found that the percentage of explained variances decreased from 33 to less than 1. With dimensions lager than 4, the value of explained variances is less than 0.1, therefore take the first 4 factors as our principal component. The Correlation plot in Figure 2 also shows that more factors didn’t seem to give more information since the color of these factors faded away as the dimension increases.

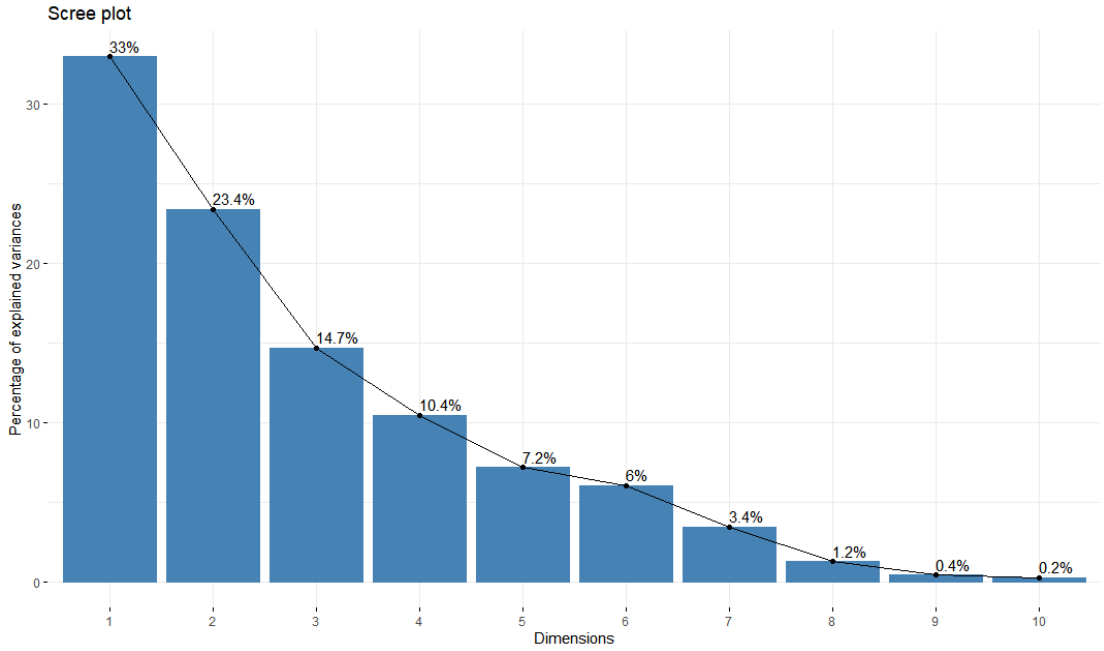

Fig. 1 PCA Scree Plot

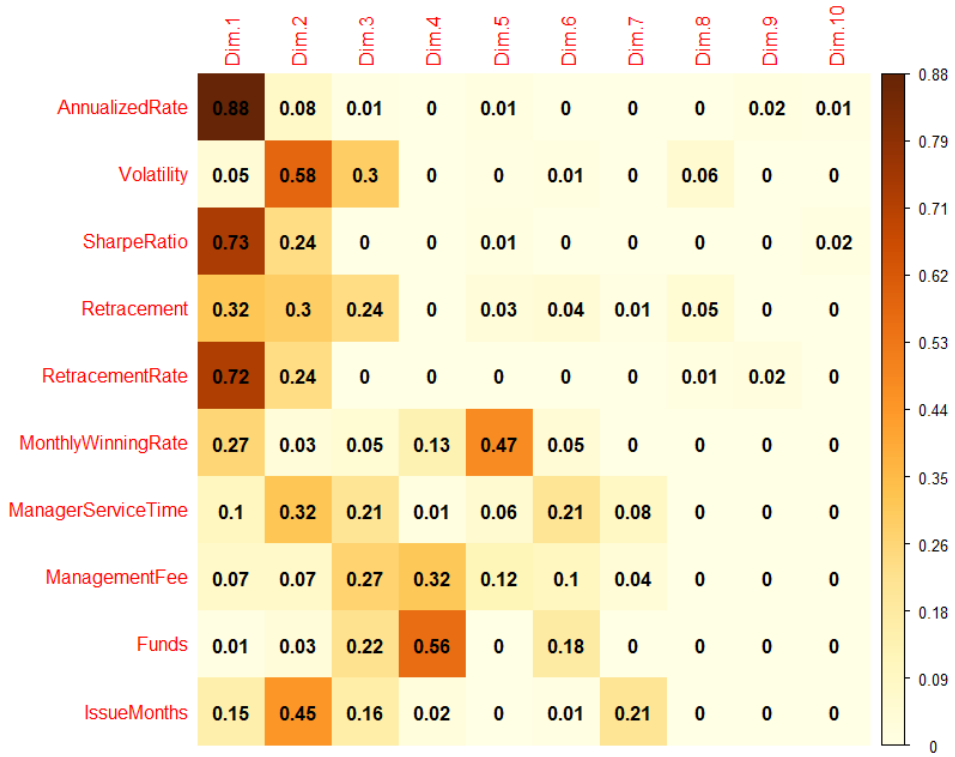

Fig. 2 Correlation Plot of Variables

Hence, with the cumulative explain of variance in table 2, four principal components can be recognized with table 3. In the rotated component matrix, variables which has absolute value more than 0.4 would conclude in that component. In this way, we have annualized rate, Sharpe Ratio and the retracement rate as the first component, mostly knowing as the income factor. Volatility, monthly winning rate and retracement are vital numbers in presenting the anti-risk ability to investors. Higher stock price volatility often means higher risk and helps an investor to estimate the fluctuations that may happen in the future. Meanwhile, high value retracement also limits the investor's risk. The monthly winning rate might come a little bit straightforward, since the ability to profit every month do reflects that they could handle the risk well enough. Management fee ratio, manager service time and ETF issue time all makes influences on the operation of funds, guaranteed the management ability

to all investors in the market. Even though there are already three main factors, the funds amount could also recognize as the volume component in principle components.

Table 2. Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.296	32.959	32.959	3.296	32.959	32.959
2	2.338	23.378	56.337	2.338	23.378	56.337
3	1.469	14.691	71.028	1.469	14.691	71.028
4	1.046	10.456	81.484	1.046	10.456	81.484
5	.717	7.171	88.655			
6	.603	6.032	94.688			
7	.343	3.434	98.121			
8	.125	1.247	99.368			
9	.040	.395	99.763			
10	.024	.237	100.000			

When receiving these factors, we assume that the corresponding ETFs are largely represent by them. To explore further differences between ETFs, taking the principal components as cluster variables and try summarizing the patterns between different ETF categories.

Table 3. Rotated Component Matrix

	Component			
	1	2	3	4
AnnualizedRate	0.951	0.224	0.095	-0.001
Volatility	0.196	-0.930	-0.125	0.055
SharpeRatio	0.985	-0.035	0.044	-0.001
Retracement	0.211	0.886	0.165	-0.065
RetracementRate	0.981	-0.026	0.020	0.011
MonthlyWinningRate	0.357	0.493	-0.065	0.326
ManagerServiceTime	0.002	0.206	0.639	0.442
ManagementFee	0.110	-0.114	0.820	-0.198
Funds	-0.014	-0.079	0.086	0.898
IssueMonths	0.009	0.309	0.800	0.212

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

As shown in Figure 3, the class spacing is gradually widened due to the condensation of classes. The class spacing increased significantly before the clustering of 9 classes, as the shape of a peak, but after 7 classes, the class spacing decreased rapidly. According to the principle that the smaller the distance of classes, the higher the degree of similarity, this study believes that it is appropriate to cluster the above indicators into 7 categories.

Table 4. Number of Cases in each Cluster

Cluster	1	87.000
	2	6.000
	3	41.000
	4	178.000
	5	1.000
	6	154.000
	7	161.000
Valid		628.000

Among clustering molecules, the average factor score of the fifth type of fund is 4.2066, which is the first ladder ETF, which mainly includes GF CSI All-Index information Technology ETF and the Southern CSI 500 Information technology ETFs, YiFangda CSI 300 non-bank ETFs, its anti-risk ability to contribute the most. Second, sixth and seventh types of ETF funds are the second tier ETFs, whose management capabilities contribute the most, with the first tier and the second tier ETF accounting for a total of 12% of the sample size, and the first tier classified as the third tier ETF, accounting for 50% of the sample data. This phenomenon is in line with the objective law that cutting-edge products only account for a small number of market quantities.

It can be seen from the F scores of seven types of funds that the performance of ETF fund products in the market is quite different, and the reason for this phenomenon may be because of the increasing diversification of tracking index targets, which makes the performance of ETF funds in China differentiated.

In addition, from the study of cluster analysis, it will be found that fund management ability will also have a greater impact on its performance, but it is not significantly reflected in the factor score. This is explained in investors with more professional qualities in the market will get better returns than ordinary investors, whether this violates the idea of passive management of ETFs needs to be repeatedly verified in the following research

4. Conclusion

According to previous studies by scholars, China's Funds market is a weakly efficient market, which gives fund products an extraordinary opportunity to make further expansion. ETF funds have the advantages of convenient portfolio risk diversification. With the characteristics of stable income, it is an ideal passive management tool, which is in line with the requirements of the country's Funds market and investment environment. Compared with other investment tools, ETF funds limit the opportunity cost for investors in a certain way. Judging from the current excess returns of ETF funds, in ETF. The trade order flexibility of ETF also gives investors the benefit of making timely investment decisions and placing orders in a variety of ways. Judging from the current excess returns of ETF funds, in the long-term investment activities of funds, investors would be able to obtain the average market return in China’s Funds market.

This research is based on the factorial analysis model constructed by 10 variables. After the processing of principal component analysis, four types of factors are obtained. Concluded as the income factor, anti-risk factor, management ability factor and the volume factor. In the descriptive statistical analysis of timing ability indicators, it is found that 45.95% of the funds have negative timing ability and do not have timing ability. In the process of clustering sample funds, it is concluded that the excellent fund types only account for 12% of the total sample number, and most ETF funds have great anti risk ability. It shows that ETF is an excellent investment tool of risk diversification, and investors can improve their anti-risk ability of total assets by purchasing it.

To sum up, fund investment companies can learn from the managers of foreign mature markets. To improve the professional quality of fund managers in combination with their own investment skills.

It is conducive to the stable and healthy development of the fund market. At present, the growth environment of ETF funds in China is good. However, the continuous emergence of innovative products in the market has led to the accumulation of investment risks. Regulators should strengthen the control of the whole market and strictly control the listing requirement of ETF products. At the same time, the market should standardize fund operations, optimize the delisting and elimination mechanism, and popularize investment knowledge to individual investors to guide then to make rational investment, optimize fund products constantly and improve investment efficiency.

In this research, we only discuss the performance of ETF funds in the whole market. To give a general direction for the construction of its evaluation system, and to a certain extent investor provide a reference for investing in ETF funds. In the constantly changing market environment, the construction of ETF fund performance evaluation model needs further research.

References

[1] Treynor J L. How to Rate Management of Investment Funds. Harvard Business, 1965, vol. 43(1): 63-75

[2] Sharpe Willian F. Mutual Fund Performance. Journal of Business ,1966, vol. 39: 119-138

[3] Michael C Jensen. The performance of mutual funds in the period 1945-1964. The Journal of Finance, 1968, vol. 23(2): 389-416.

[4] Kothari S P, Jerold B. Warner. Evaluating Mutual Fund Performance. The Journal of Finance, 2004, 56(5):1985-2010.

[5] Dellva, W L, Olson, G T. The Relationship between Mutual Fund Fees and Expenses and Their Effects on Performance. The Financial Review, 1998, 33(1): 85-103.

[6] Mann Thomas, Atra Robert J, Dowen Richard. U.S. monetary policy indicators and international stock returns: 1970-2001. International Review of Financial Analysis, 2004, vol. 13(4): 543-558.

[7] Chen Yongsheng, Yang Ning. Do Investors Have the Ability to Choose Funds Correctly. Macroeconomic Research, 2011(05):15-24.

[8] Qin Zhengyan. An Analysis of the Investment Value of Listed Companies in China's Financial Industry—Based on the Financial Index System. Business Economics, 2020(1):88-89.

[9] Zou Jing. Research on Internet Enterprise Value Evaluation—Based on the Case Analysis of Tencent Holdings. Southwestern University of Finance and Economics, 2016(9):1112-1120

[10] Zhang Tiancheng. Analysis of Influencing Factors of Shanghai Stock Exchange 180 Stock Return. Kunming: Yunnan University of Finance and Economics, 2019.

[11] Shi Biao, Xia Liyu, Yu Xiahua, Wang Yan. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[12] Girish Punj, David W. Stewart. Cluster analysis in marketing research: review and suggestions for application, 1983, 20(2):134–148.