INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES
DOI: 10.46300/9101.2020.14.15 Volume 14, 2020

Arbitrage Pricing Model Based on Factor Analysis-Random Forest Regression and its application

Xiaoyang Zheng[1]*, Guyue Tian[2], Fengsi Yu[2]
[1]Institute of Liangjiang Artificial Intelligence, Chongqing University of Technology, Chongqing, 400054, China
[2]College of Science, Chongqing University of Technology, Chongqing, 400054, China

Received: January 16, 2020. Revised: March 27, 2020. 2nd Revised: April 28, 2020.
Accepted: June 18, 2020. Published: July 20, 2020.

Abstract—Firstly, this paper establishes K-factor linear model and arbitrage pricing model (ATP) according to 'the Asset Pricing Model-Arbitrage Pricing Theory', Then from 2001 to 2017, the Statistical Yearbook of the National Bureau of Statistics collected 10 factors as the original factors such as gross national product, gross industrial product and gross tertiary industry product. After synthesis and simplification, three common factors are extracted to replace ten original factors.The first common factor variable is used to reflect the overall economic level of the country;The second common factor variable reflects a country's inflation rate;The third public factor variable reflects the total annual net export trade situation of the country.

After the common factor is determined, the value of the common factor is calculated from the original data.Collect the annual return of 10 stocks for 17 years and do twice random forest regression,we get the arbitrage pricing model. Then, based on the same common factor data, another arbitrage pricing model is obtained by imitating the linear regression method of previous similar papers. By comparing the pricing error, we can find the pricing effect of the model obtained by random forest regression is better than that of the model obtained by linear regression.

Keywords—Arbitrage Pricing Model, Factor Analysis, Random Forest Regression,KMO Test

I. Introduction

The efficiency of capital and resource allocation is one of the main contents of financial economics. In the market economy, the allocation of resources is realized through the trading behavior of a large number of market participants, who not only determine the supply and demand of capital, but also determine the price fluctuation, and the price mechanism leads to the supply and demand of capital and its final allocation.Therefore, the pricing mechanism, the behavior of market participants and the market environment have jointly stimulated the development of the capital market, which is called 'Troika'[3].

This paper analyzes and discusses the arbitrage pricing model in the capital pricing of the 'Troika'.In the theory of financial economics, there are two main methods to determine the current reasonable value of a risky asset that is assumed to be a known distribution of future returns: one is the pricing method based on competitive equilibrium theory, such as the capital asset pricing model developed on the basis of portfolio theory;The other is the pricing method based on arbitrage pricing theory (APT), which was put forward by Ross in 1976, this model shows that the return on capital assets is the result of various factors, such as inflation, stock market composite index, industrial growth index and so on[1].

In China, the research and application of arbitrage pricing model is very extensive. For example, Ding[10] uses this model to analysis the Shanghai Stock 50 index. Guo[12] studies the applicability of apt theory in China's securities market, and

made an empirical analysis with the financial industry as an example. Li[13] studies the factors influencing stock price volatility according to APT model. Yang[14] analyzes the efficiency of China's securities market during the financial crisis based on apt theory. However, in the existing empirical research, the analysis methods adopted by the authors are often of large error, and there is room to improve the goodness of fit of the model. Therefore, this paper adopts a new method to build the model, trying to reduce the error and improve the goodness of fit.

II. Arbitrage pricing model and factor analysis-random forest regression method

In order to make the model reasonable, we assume that investors have the same investment philosophy, the market is complete and investors are risk-averse and can maximize utility.

2.1 Establishment of Arbitrage Pricing Model

In 1976, Stephen Ross, an American scholar, published a classic paper 'Arbitrage Theory of Capital Asset Pricing' in 'the Journal of Economic Theory', and put forward a new asset pricing model, namely Arbitrage Pricing Theory (APT Theory)[1].

This theory uses arbitrage concept to define equilibrium, does not need the existence of market portfolio, and requires fewer and more reasonable assumptions than the capital asset pricing model (CAPM model).The theory assumes that the rate of return of any risky security is affected by k factors.From a k-factor linear model, the following formulas can be given:

$$\tilde{r}_i = a_i + \sum_{k=1}^{k} b_{ik}\tilde{f}_k + \tilde{\varepsilon}_i (i = 1,2,...,n) \qquad (1)$$

In formula (1), there are:

$$E(\tilde{\varepsilon}_i) = E(\tilde{f}_k) = E(\tilde{\varepsilon}_i \tilde{\varepsilon}_j) = E(\tilde{\varepsilon}_i \tilde{f}_k) = 0 \qquad (2)$$

$$E(\tilde{\varepsilon}_i^2) = s_i^2 < S^2 \qquad (3)$$

$\tilde{r}_i$ represents the rate of return of the ith risky security, a_i represents the average return of a risky security i when all the factors affecting the return of the risky security are zero; $\tilde{f}_k$ represents the value of factor k-th; b_{ik} represents the sensitivity of the risky security i to factor k-th; $\tilde{e}_i$ represents a random perturbation term;

When there is no gradual arbitrage opportunity, the following approximate pricing model, namely arbitrage pricing model (APT), can be obtained from k-factor linear model:

$$E(\tilde{r}_i) = a_i \approx \lambda_0 + \sum_{k=1}^{k} b_{ik}\lambda_k \qquad (4)$$

Note the error of the above formula as $v_i = a_i - \lambda_0 - \sum_{k=1}^{k} b_{ik}\lambda_k$;When there is no progressive arbitrage opportunity in the market, we can get $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} v_i^2 = 0$, λ_k represents the risk premium on risk security i for factor k-th.

The key to the establishment of arbitrage pricing model lies in the screening of factors.However, the return rate of a risky security is affected by many factors, so we need to determine the number of factors to build APT model.Assuming that there are n factors that affect the return rate of securities, there may be a combination of $\sum_{m=1}^{n} C_m^n$ factors. In order to simplify the calculation steps and quickly find the optimal model, we adopt factor analysis method to construct and screen factors.

2.2 Factor Analysis and Random Forest Regression

Factor analysis is a statistical technique for extracting common factors from many variables, which was first proposed by C. E. Spearman, a British psychologist.Finding representative factors hidden in many variables and reducing the same essential variables to one factor can not only reduce the number of variables, but also test the hypothesis of the relationship between variables.

Generally speaking, there are two basic methods to identify and determine factors: statistical method and reasoning method.Statistical methods involve determining factors from a comprehensive set of asset returns and using sample data from those returns to construct a portfolio of assets representing the factors.For example, ConnorKorajcxyk (1988) [4], Lehmnn and Modeu (1988) [5], the former used factor analysis method, the latter used principal component analysis method.Reasoning methods are based on the systematic risk principle of capture economy to identify factors.

In the arbitrage pricing model, the risk of stock return can be divided into two parts: the dispersible risk and the non-dispersible risk. The mean value of the dispersible risk is zero, which can be ignored in a large sample. The non-dispersible risk part is determined by K common factors, and can reflect the relationship between stock return and each non-zero risk profit through the coefficient of k factors.Each common factor can represent one to many different indicators in the national economy.To sum up, this paper introduces factor analysis method to screen the factors of APT.

After K common factors are screened out, the arbitrage pricing model is fitted by means of random forest regression;Random forest is an ensemble algorithm, which combines multiple weak classifiers, and the final results are voted or averaged, so that the results of the overall model have high accuracy and generalization performance.The specific steps are as follows:

Step 1: select m samples from the original training set, perform n _ tree sampling, generate n _ tree training sets and train n _ tree decision tree models respectively.

Step 2: Suppose that the number of training samples is n, the best feature is selected according to the Gini index to split each time, and the split continues until all training samples of the node belong to the same class.

Step 3, forming a random forest by generating a plurality of decision trees. For the regression problem, the final prediction result is determined by the mean value of the prediction value of multiple trees.

Because of the simple realization, high precision and strong anti-overfitting ability of random forest regression, this paper uses random forest regression to fit the arbitrage pricing model.

III. Collection and Test of Original Index of Factor Analysis

3.1 Data collection

In the previous literature[6-9], economists generally believe that there are at least three different factors in APT arbitrage pricing model: indicators of overall economic activity, inflation rate and some types of interest rate factors.In view of this, this paper collects the gross national product, gross industrial product, gross secondary industry product, gross tertiary industry product, national consumption level, inflation rate, total fixed assets investment, total retail sales of consumer goods, total money supply,The 10 factors of the total annual net export trade are used as the original index variables for factor analysis, and the collected data are shown in the following figure 1.

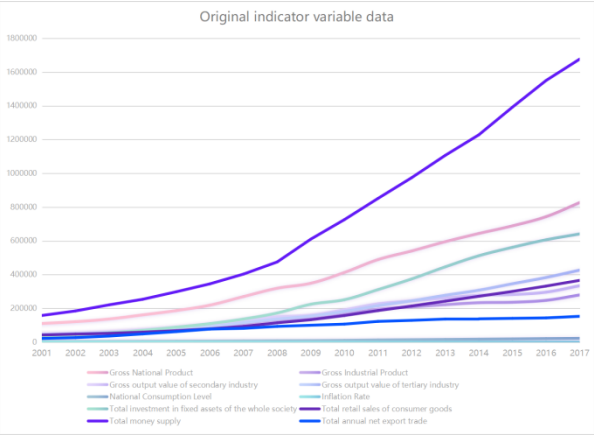

Figure 1. Original indicator variable data

3.2 Data validation

Because factor analysis is to construct a few representative common factor variables from many original index variables, which has a high requirement for the correlation between the original index variables, so it is necessary to analyze the correlation of the original index variables before factor analysis.

In this paper, the KMO (Kaiser-Meyer-Olkin) test is used to collect the original index variables.The result shows that the value of KMO is 0.771, which is in line with the standard proposed by Kaiser[9], so the original index variable is suitable for factor analysis.

IV. Using Factor Analysis to Determine the Combination of Factors in APT

Factor Analysis is performed on 10 original index variables to obtain the variance and cumulative values of each component as shown in Table 1 below.

Table 1 Variance and Accumulation of Components

Total variance of interpretation				
Compositions	Initial eigenvalue		Extracting the sum of squares loading	
	Percentage of variance	Cumulative percentage	Percentage of variance	Cumulative percentage
1	89.155	89.155	89.155	89.155
2	7.460	96.616	7.460	96.616
3	3.258	99.873	3.258	99.873
4	0.084	99.957		
5	0.032	99.989		
6	0.007	99.996		
7	0.003	99.999		
8	0.001	100.000		
9	5.11E-05	100.000		

It can be seen that the cumulative variance contribution rate of the first three principal components has reached 99.873%, which conforms to the extraction principle of the main components, so the first three principal components are extracted and the component matrix corresponding to the principal component is shown in Table 2.

Table 2 Composition Matrix

Component Matrix			
	Compositions		
	1	2	3
Gross National Product	0.999	-0.032	0.007
Gross Industrial Product	0.993	-0.006	-0.106
Gross output value of secondary industry	0.995	-0.011	-0.086
Gross output value of tertiary industry	0.995	-0.05	0.086
National Consumption Level	0.996	-0.047	0.067

Inflation Rate	0.263	0.964	0.034
Total investment in fixed assets of the whole society	0.993	-0.049	0.088
Total retail sales of consumer goods	0.994	-0.054	0.094
Total money supply	0.994	-0.052	0.095
Total annual net export trade	0.062	0.049	-0.926

As can be seen from the composition matrix in Table 2, except the inflation rate and the total annual net export trade, the other eight index variables, such as GNP and GDP, are more effective to the load of the first factor variable, reflecting the information of the overall economic level of the country. Similarly, it can be seen from Table 2 that inflation rate is more effective for the load of the second factor variable, so it can be considered that the second factor variable reflects the relevant information of inflation rate, and the third variable load of annual net export trade total has great explanatory power, which reflects the relevant information of national annual net export trade total.so we can draw a conclusion: APT model is mainly related to the country's overall economic level, inflation rate and international trade.

Through Table 2, the calculation formulas of the three common factors are as follows:

$$F_1 = 0.999x_1 + 0.993x_2 + 0.995x_3 + 0.995x_4 + 0.996x_5 + 0.263x_6 + 0.993x_7 + 0.994x_8 + 0.994x_9 + 0.062x_{10} \quad (5)$$

$$F_2 = -0.032x_1 - 0.006x_2 - 0.011x_3 - 0.05x_4 - 0.047x_5 + 0.964x_6 - 0.049x_7 - 0.054x_8 - 0.052x_9 + 0.049x_{10} \quad (6)$$

$$F_3 = 0.007x_1 - 0.106x_2 - 0.086x_3 + 0.086x_4 + 0.067x_5 + 0.034x_6 + 0.088x_7 + 0.094x_8 + 0.095x_9 - 0.926x_{10} \quad (7)$$

In the formulas, x_1 : Gross National Product, x_2 : Gross Industrial Product, x_3 : Gross output value of secondary industry, x_4 : Gross output value of tertiary industry, x_5 : National Consumption Level, x_6 : Inflation Rate, x_7 : Total investment in fixed assets of the whole society, x_8 : Total retail sales of consumer goods, x_9 : Total money supply, x_{10} : Total annual net export trade.

Then,we calculated the value of F_1, F_2 and F_3 based on the data collected from the original variable from 2001 to 2017.

V. The Structure of Arbitrage Pricing Model Based on Random Forest Regression

In the past similar papers, most of the fitting arbitrage pricing models are linear regression such as J.M. Sun[7]. However, as Sun mentioned in the paper, "when the sample size is too small, the linear regression may lead to the low significance and goodness of fit of the regression equation, and eventually lead to a large error in the prediction results."[7]. Therefore, this paper innovatively uses random forest regression to fit the arbitrage pricing model.

According to the formula (1), the annual returns of 10 stocks from 2001 to 2017 are selected as the explanatory variables, which are Fiyada A, China Agricultural Science and technology, Nanbo A, Ping an bank, new good, Shahe shares, Shenchiwan A, Shenhua A Shenzhen KonkaA and Shenzhen Science and technology.Then the three common factor variables corresponding to a total of 16 years of data as explanatory variables for random forest regression, get the value of each stock a_i 、 $b_{ik}(k=1,2,3)$.

Then according to the formula (4), with a_i as the explanatory variable and $b_{ik}(k=1,2,3)$ as the explanatory variable, the random forest regression fitting is used again to obtain the arbitrage pricing model (recorded as $APTM_1$):

$$a_i = 0.164 - 1.06\lambda_1 + 0.022\lambda_2 - 0.37\lambda_3 \quad (8)$$

Because APT model is an approximate pricing model, so when applied to individual stocks, there may be a large error, so this paper uses the above ten stocks to construct a simple equal-weight portfolio, according to $\sigma = \frac{1}{n}\sum_{i=1}^{n}(\bar{r}_{i\text{predicted value}} - \bar{r}_{i\text{True value}})^2$ to measure the error.At the same time, in order to make a comparison with the linear regression method , we select the above ten stocks to get an arbitrage pricing model by linear regression method and record it as $APTM_2$, and calculate the pricing error of this model, we can get the pricing error of the two regression methods as follows:

model	$APTM_1$	$APTM_2$
Pricing Error	0.00836	0.12935

The pricing error $\sigma = 0.00836$ of the model $APTM_1$ is much smaller than that of the linear regression model. Therefore, it can be concluded that the random forest regression method is better and the model has better pricing effect.

From the formula (8), it can be found that the sensitivity coefficient of risk securities i to inflation rate is positive, which means that when the risk premium of risk securities i to inflation rate is larger, the expected return of the securities will be larger; On the other hand, the sensitivity of risk securities i to the national total economic level and the total fixed assets investment of the whole society is negative, which shows that when the risk premium of risk securities i to inflation rate is larger, the expected return of the securities will be smaller.

VI. Conclusion

In this paper, the factor analysis method is introduced to synthesize and simplify the data of 10 original indexes, such as GNP, GDP, inflation rate and total annual net export.According to the results of factor analysis, this paper extracts three main public factors, referring to literatures[11-14], we can think that these three public factors represent the

overall economic level, inflation rate and total fixed asset investment of the whole society.

On this basis, this paper creatively uses random forest regression to establish the corresponding arbitrage pricing model. In order to verify the quality of the model, this paper uses real-time data to test, and compared with the traditional linear regression method, the smaller error term shows that the effect of random forest regression is better. Therefore, the method proposed in this paper is better than the traditional method in establishing the arbitrage pricing model, and it is easy to apply and promote.

Furthermore, this paper believes that if we expand the sample size and select more representative stocks as the sample, we can further reduce the pricing error of APT.

References

[1] R. Stephen, Arbitrage theory of capital asset pricing. Journal of Economic Theory, 1976.
[2] L.Y. Cai, Quantitative investment using Python as a tool. Electronic Industry Press, vol. ED-23, pp. 132-138, 2017.
[3] Y. Zhang, Empirical Test of Interest Pricing Theory in Shanghai Stock Market. Journal of World Economics, 2000.
[4] G. CONNOR, & R. KORAJCZYK, Risk and retum in an equilibrium APT: application of a new test methodology. Journal of Financial Economics, vol. ED-21, pp. 255-290, 1998.
[5] E. FAMA, & K. FRENCH, Dividend yields and expected stock retums. Journal of Fïnancial Economics, vol. ED-22, pp.3-27, 1988.
[6] E. FAMA, & K. FRENCH, Multifactor explanations of asset pricing anomalies. Joumal of Fïnance, pp.51-55, 1996.
[7] J.M. Sun, An Arbitrage Pricing Model Based on Factor Analysis and Its Empirical Study. Journal of Finance and Trade Research, 2007.
[8] J.K. Liu, and D.H. Mao, Application of Factor Analysis in Arbitrage Pricing Model. Journal of Shandong Agricultural Administrators' College, 2013.
[9] J.P. Zhu, Application of multivariate statistical analysis. Science Publishing House, vol. ED-7, pp. 58-69, 2010.
[10] M.Y. Ding, Y.M. Pan, and Y.Q. Ding, Empirical Test of Arbitrage Pricing Model for the SSE 50 Index Stocks. 2018 International Conference on Computer Science, Electronics and Communication Engineering, 2018.
[11] Miklós Rásonyi, Maximizing expected utility in the Arbitrage Pricing Model. Journal of Mathematical Analysis and Applications, 2017.
[12] S.M. Guo, The Applicability of APT Theory in China's Securities Market--Taking the Financial Industry as an Example. Journal of Guangxi Quality Supervision Report, pp. 12-13, 2019.
[13] S.P. Li, Research on the factors influencing stock returns under the arbitrage pricing model. Journal of Chinese market, pp. 27-29, 2019.
[14] Z.Y. Yang, and Z. Cao, An Analysis of the Efficiency of China's Securities Market in the Post-Financial Crisis Era Based on Arbitrage Pricing Theory. Journal of Commercial Economics, pp. 18-20, 2018.