Empirical Research of Energy Sector Based on Principal Component Analysis

Haoran Sun, Jie Li*, Feiyu Long

Fuzhou University, Fuzhou, Fujian, 350108

E-mail: vj162774@163.com

Keywords: single factor test; principal component analysis; return and risk prediction; multi-factor model

Abstract: In this paper, 76 stocks in the energy sector were selected as sample stocks and data from 2013 to 2017 were used for single factor test to select pre-selected factors, factor orthogonality was carried out through principal component analysis, return prediction was made by principal component factors, and quadratic optimization considering risks was introduced to study how the multi-factor model could guide investors' asset allocation in the energy sector. In the empirical, without introducing the future data and considering new listings, the empirical parameters were obtained by solving the model with in-sample data to realize the weight proportion of the selected stocks. It obtained an annual yield of 46% out of the sample with a high winning rate, and beat the HS300 index with the energy sector with a low valuation, highlighting the weight matching ability of the strategy and verifying the effectiveness of the model. This model uses principal component analysis to reduce the noise of factor information, and combines risk analysis to optimize the feasibility of the strategy. It provides a new perspective for multi-factor quantitative investment.

1. Introduction

The energy industry, as a pillar industry of the country, has many stocks with low valuations and investment value. Therefore, how to obtain excess returns in the energy sector investment has research value [1]. Currently, the more mainstream method in quantitative investment is multi-factor stock selection. This model is mostly used in the a-share market for asset allocation. Many investors and scholars' research and improvement of the multi-factor model have also confirmed the effectiveness of the multi-factor stock selection model in quantitative investment [2].

2. Single factor test

2.1 Purpose of Single Factor Testing

In order to build a multi-factor stock selection model, historical data is required for single-factor backtesting. By testing the factor stock selection ability and income interpretation ability, we can filter out the factors that are more effective during the sample period and use them as the foundation of the out-of-sample multi-factors stock selection model.

2.2 Data Acquisition

The data source is Shanghai Shangyang Asset Management Co., Ltd. The data in the sample is 76 stocks in the energy industry selected from 2013 to 2016 for a total of 971 days, and the data outside the sample is a total of 610 days in 2017-July 2019 [3]. Distinguishing the data from the inside and outside the sample can effectively avoid the model error caused by the introduction of future data. There are seven in the factor library. There are 91 major factors, with a total of 91 small factors. The major factors are Fluidity, Growth, Momentum, Size, Technique, Value, Volatility.

Table 1 76 stocks in the sample

Stock code									
sz000059	sz000937	sh603689	sh601857	sz002128	sz002554	sz300191	sh600339	sh600508	sh600777
sz000159	sz000968	sh603619	sh601808	sz002207	sz002629	sh600028	sh600348	sh600546	sh600792
sz000552	sz000983	sh603113	sh601798	sz002221	sz002778	sh600121	sh600387	sh600583	sh600856
sz000554	sh603906	sh603036	sh601699	sz002267	sz002828	sh600123	sh600395	sh600688	sh601101
sz000723	sh603800	sh603003	sh601666	sz002278	sz300084	sh600157	sh600397	sh600725	sh601088
sz000780	sh603727	sh601918	sh601225	sz002353	sz300157	sh600188	sh600403	sh600740	sh601015
sz000852	sh603798	sh601898	sz002018	sz002490	sz300164	sh600256	sh600408	sh600759	sh601011
sh601001	sh600997	sh600985	sh600971	sh600871	sz000096				

2.3 Data Standardization

This article uses the method of five times the standard deviation to de-extreme the centralization of all factor data, and assigns 0 to the missing factor data due to various reasons. Finally, by doing a standardization process the data of each stock in each period reduces the mean value to double the standard deviation, the above operation can reduce the impact of outliers and make subsequent single-factor tests more convenient. It is important to point out that in the process of single-factor tests, in order to avoid being caused by the listing of new shares, the fact that the word board cannot be bought, during the factor test process, we will limit the backtesting after the seventh period (the position adjustment cycle is one week and the position adjustment day is every Monday). Test the stocks after 30 days of listing. This effectively avoids the above problems and makes the selection of factors and the subsequent construction of a multi-factor model more reasonable.

2.4 Factor Test Indicators

Because the choice of factors plays a decisive role in the profitability of the multi-factor stock selection model, it is necessary to describe as many dimensions as possible during the factor test [4]. This article selects three dimensions for testing. The ability to distinguish, the factor's ability to explain stock returns, and the factor's rate of return win (secondary). Among them, the stock selection ability will be tested by grouping the stocks, and there are long and short portfolio return difference indicators and descending grouping yield curve indicators. The explanatory power of stock returns will be described by the 3-period rolling rank correlation coefficient, the rank correlation coefficient during the sample period, and the IR value calculated by the single-period rolling rank correlation coefficient. In addition, the regression test of the single factor and stock returns per cycle will provide t value, adjusting the R-squared and the factor's rate of return index. The former two are still the factor's description of the stock's ability to explain stock returns, while the latter can calculate the factor's rate of return as a separate index. These indicators are detailed below.

(1) Long-short combination indicator: The factor value is called factor exposure, and the difference between the periodic returns of the largest combination of factor exposure and the smallest combination of factor exposure is used to draw the return curve. Obviously, when the upper right of the curve is stronger, the factor's stock selection ability is stronger [5].

(2) Descending ordering return curve: Use the factor data size to sort the stocks in descending order, divide them into 5 groups, and then draw the return curves of each group of stocks separately. When the distance between the curves is larger, the degree of differentiation is more obvious when the factor is selected.

(3) Rank correlation coefficient: Spearman correlation coefficient, which is the correlation description of factor data and stock cycle returns. It is only sensitive to the ranking results and not sensitive to specific numbers. When the absolute value of the rank correlation coefficient is greater, the ability of the factor pair to explain stock returns is stronger, of course, the indicator also reflects the factor's ability to predict stock returns.

(4) Information ratio IR: In the industry, the ratio of the single-period rank rolling correlation coefficient to the standard deviation is used as the calculation of the information ratio IR. The larger

the absolute value, the more it reflects the predictive interpretation ability of the factor.

(5) Regression index t value and adjusted R-squared: Regress the data of the previous period of the factor and the return of the next period of stocks in each cycle to obtain the t-value and adjusted R-squared of each period. Among them, the t-value greater than 2 is compared with the total number of cycles. The T index is constructed. The larger the T index, the stronger the factor's ability to explain. In the industry, the adjustment of the slice time for one week is more than 0.2. If it can be close to 0.2, then the explanatory power of the factor will be very good.

(6) Factor rate of return: The factor rate of return is the cross-sectional regression coefficient of each period of factor. The prediction direction of the factor is determined by the plus and minus of the rank correlation coefficient during the sample period, and the direction of the factor rate of return is determined by the positive and negative sign decision of regression result of each period. The ratio of the period of the factor return to the rank correlation coefficient in the sample period to the total number of cycles can be used to obtain the index of the rate of win of the factor return.

2.5 Factor Test Analysis

This section selects the test results of a factor for analysis examples.

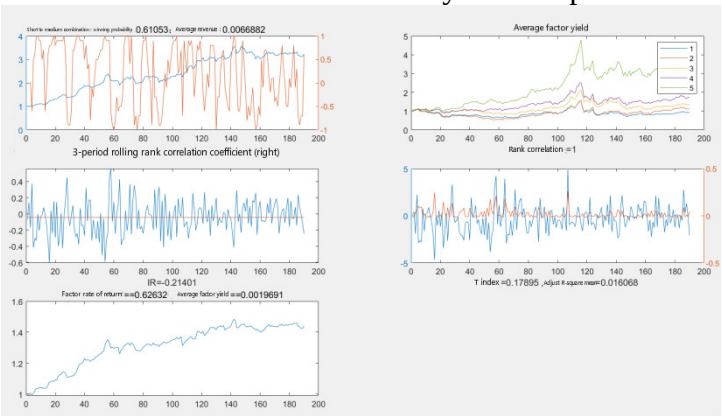

Figure 1 Factor test analysis

The test factor is SGRO_5, which belongs to the Growth category. According to the test results in the figure above, it can be seen that the long and short performance of the factor and the descending grouping performance are relatively good, of which the rank correlation even reaches the upper limit 1. In addition, the winning ratio of the factor return rate reaches 62.63. In summary, this factor can be included in the construction of a multi-factor stock selection model.

2.6 Test Results and Inclusion Factors

After testing, 24 good factors were selected:

Table 2 Selection factors

Factor type	Fluidity class	Growth class	Momentum class	Size class	Technique class	Value class	Volatility class
Inclusion factor	STOQ	SGRO_5	AlfaHS300_1m	log_Size	KDJ_D	EP_a	HSIGMAHS300_12m
	STOQ_barra	EGRO_5	log_marketcap	log_asset	KDJ_K	EPttm	
	STOS	growth_ocf_qq_3	DiffClose_ma60Zx		MACD_difdea	ncfpttm	
	STOS_barra	growth_ocf_qq_5	RSTR_m1		MACD_emadiff		
	yoy_or_qq		RSTR_m3				

Through the factor test, it can be seen that during the sample backtest period from 2013 to 2016, the overall performance of liquidity and momentum factors in the energy industry is better, while the overall performance of scale factors and volatility factors is poor.

2.7 Factor Orthogonal and Multicollinearity Problems

The multi-factor stock selection model is a linear regression model. If there is a high correlation

between the explanatory variables in the linear regression model, the model will be distorted or difficult to accurately estimate [6]. So before establishing a profit forecasting model, multiple factors need to be considered. Colinearity is processed accordingly. The colinearity of factors in the industry can be processed in the following ways:

Direct elimination: For factors that show a high correlation with other factors but cannot provide more information, direct elimination is generally used.

Factor synthesis: For small class factors with certain correlation that are not willing to be eliminated directly, multicollinearity can be solved by synthesizing small factors into large factors. Equal weights can be used when factors are combined, or other weighting methods can be used.

Factor orthogonality: Regress one of the factors with higher correlation with another factor, perform stepwise regression, and take the regression residual term instead of the factor value [7].

This article will use the principal component analysis method to perform factor orthogonalization to solve the multicollinearity problem of the model. This method is an effective data dimensionality reduction method, which can effectively solve the multicollinearity problem of variable information overlap and regression models. And finally build a multi-factor risk stock selection model.

3. Principal component analysis

3.1 Purpose of Principal Component Analysis

In a multi-factor stock selection strategy, the number of factors selected through a single factor test may still be large, and there may be complex correlations between the factors, thereby increasing the complexity of the problem analysis. The principal component analysis method is reducing regression. At the same time as the number of factors, you can minimize the useless information contained in the original data to achieve the best use of the data. In addition, the algorithm can also orthogonalize the factors, effectively solving the variable information overlap and multicollinearity linearity regression problem. Finally, the factors obtained through principal component analysis will be used as variables in the multi-factor return prediction model.

It is important to point out that because the number of factors in the factor pool is greater than the number of stocks, in the principal component analysis algorithm, in order to prevent the singularity of the correlation coefficient matrix of the factors, some factors must be selected during the single factor test, and the total number is less than the number of stocks. This can ensure the smooth running of the principal component analysis algorithm.

3.2 Introduction to Principal Component Analysis

The principal component analysis method is the most widely used data dimensionality reduction algorithm that can denoise a large amount of data. The principal component direction of each dimension is selected and based on the orthogonalization, the ratio of the original information data is mapped from n-dimensional space to k-dimensional space. The set of orthogonal bases mapped on k-dimensional space is the principal component that is finally constructed.

In actual research, because the purpose of the principal component is to reduce the dimension and reduce the number of variables, a small number of principal components are generally selected. The general index of the industry is about 20, and this article is set to 18 according to the test results. The internal average can cover 90% of the original data information.

This article uses the eigenvalue factor screening method for principal component analysis and value calculation of $c_1,\cdots,c_n$, as follows:

With m indicator vectors $x_1, x_2,\cdots, x_m$, its value in the k-th experiment $a_{k1}, a_{k2},\cdots, a_{km}, k=1,2,\cdots,n,$ write them in matrix form as

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

The matrix a is called the design matrix.

$c_1,\cdots,c_n$ is selected as the feature vector of matrix A^TA. The principal component factor is selected from the selected corresponding feature value. The weight of the data information retention is determined by the sum of the selected feature value and the sum of all feature values.

3.3 Principal Component Analysis Factor Results and Revenue Forecast

According to the industry multi-factor model experience and repeated parameter adjustment experiments, it is finally determined that the main component parameters of each multi-factor model are 18, and the parameters can cover 90% of the data information, thereby achieving the purpose of noise reduction and data dimensionality reduction.

The multi-factor return forecasting model is to use the factors obtained by the principal component analysis as variables of the multi-factor regression model to perform return forecasting. Among them, the calculation of the return forecast value of each period uses the factor return rate and the current factor returned in the previous period. The product of the values solves the problem of introducing future data.

However, the shortcoming is that the economic meaning of factors is missing due to the orthogonalization method of the algorithm, which makes it impossible to perform further factor economic meaning analysis from the obtained principal component data, and it cannot be combined with the overall macroeconomic environment to demonstrate.

4. Multi-factor risk model stock selection strategy

4.1 Purpose of Building a Multi-Factor Risk Model

The multi-factor income prediction model often needs to consider risk factors in practical application, because the income prediction model may be over-exposed to factors with large volatility in returns during stock selection, which will cause greater risk in the finally selected stock portfolio. Therefore, we also need to introduce a risk prediction model into the model to control the possible retracement of returns. We also use the risk coefficient that describes the risk to optimize the model's returns in each cycle, and finally give a multi-factor stock selection model and related parameter value.

4.2 Risk Secondary Planning Model

An effective way to introduce risk for asset allocation is the risk quadratic programming model. Traditionally, the risk covariance matrix of the selected stock is often used to measure the risk. This method often requires a large number of matrixes under the premise that the matrix is invertible. Therefore, according to the Barra risk model, this article chooses to characterize the future volatility of the stock pool by estimating the covariance matrix of the factors obtained after the principal component analysis, that is, the volatility of the factors is used to measure the market risk. The larger the variance matrix, the greater the risk. In the risk model, the covariance matrix is required to be invertible. The Barra risk model can greatly reduce the number of parameters to be estimated, and it can also reduce the errors caused by the large amount of data needed to estimate the return.

Therefore, the multi-factor model is composed of two parts: profit forecast and risk forecast. The final multi-factor risk model is as follows:

$$\max \quad \alpha' w - \frac{1}{2}\lambda w' C w \quad (1)$$

$$s.t. \quad 0 \le w_t \le k_t \quad (2)$$

$$1'w = 1 \quad (3)$$

Where w is the weight to be configured for the selected stock, (1) is an optimization objective function that considers risk, and (2), (3) are conditional constraints.

(1) $\alpha'w$ is the combined income forecast for each period, its calculation uses the product of the factor return rate returned in the previous period and the factor value of the current period; $w'Cw$ is the combined risk forecast, where the Barra risk model is used by C to convert the risk of individual stocks into the sum of the multi-factor model risk and the residual risk of the stock itself: $C = X_f F X_f' + \Delta$, where X_f is the factor exposure matrix for the stocks in the portfolio, F is the covariance matrix between factor returns, Δ A diagonal matrix composed of the volatility of individual stock residuals. λ is the risk aversion coefficient, the greater the value, the greater the degree of risk aversion. In the model, set the range of λ [0.1,5], with a step size of 0.1 to ensure the impact of risk factors on the investment strategy;

(2) The second condition is the restriction on the weight of each stock. Considering the actual situation that a share is not allowed to be short, the minimum weight of each stock is set to 0;

(3) The third condition restricts the sum of weights to 1.

According to the above model, the risk-recovery forecast is obtained and visualized. After multiple parameter adjustment experiments, the optimal multi-factor stock selection model parameters can be obtained, and further testing of the out-of-sample data can be performed to verify the model's performance.

4.3 2013-2017 Backtesting of Stock Strategy Samples (Compared To CSI 300 Index Returns)

During the backtest in the sample, the allocation of funds was divided into two parts, one was investing in stocks, and the other was investing in currency funds with an annual profit of 4%. The positions were adjusted weekly. In each cycle, the strategy's risk aversion coefficient was (0.1 5), take 0.1 as the optimal step size, that is, the investor's risk aversion coefficient is determined by the value of the optimal secondary planning per cycle. During the backtesting process, it was found most of the λ values are taken as 0.1, indicating that most of the forecasts before each cycle believe that the next cycle of investment in stocks will have a return, and all funds should be invested in stocks, but there are also cycles in which all funds are invested in currency funds, and some funds hold positions in stocks almost none.

During the adjustment of the strategy parameters, it was found that when the predicted return of the subsequent period is> 0.12, the borrowed funds are allowed to increase leverage and set the leverage to three times; when the predicted return of the subsequent period is> 0.1, no leverage is added; when the predicted return is less than 0.05, no leverage is added; otherwise, all the money funds are invested. It is important to point out that in the investment process, the part of the forecasted return of the next period <0 is also invested in stocks, which indicates that the stock strategy has a certain degree of reverse operability. In the case of increasing leverage, it is not ruled out that all funds are ultimately invested in currency funds, because domestic stocks cannot be shorted, and the weights of the secondary optimal solution may all be 0. In addition, the above strategy range and parameters are test experience values, and the strategy will also use this parameter to invest in 2017-July 2019. The following figure shows the results of backtesting in the sample, the blue curve is the CSI 300 return curve, and the red is the strategic return, the pink is the return curve of the money fund. It can be seen that the backtest results are better, and the strategy is very sensitive to the capture of the rally.

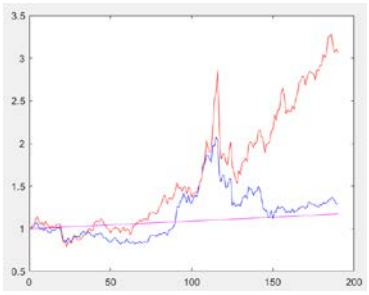

Figure 2 In-testing of stock strategy samples in 2013-2017

4.4 2017-2019 Stock Strategy Out-of-Sales Return Results (Compared to CSI 300 Index Returns)

Test the strategy out-of-sample. The test results are shown in Figure 3. It can be seen that the success of the multi-factor stock selection model has finally yielded an annualized return of 46%, and the winning rate is almost 100%.

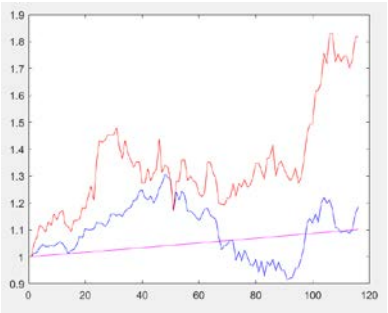

Figure 3 External backtesting of 2017-2019 stock strategy samples

4.5 Analysis of Profit Result and Deficiencies of Stock Selection Strategy

Considering that style rotation may occur in the a-share market, the multi-factor strategy should be reset periodically to obtain a stable high return. In the strategy, empirical reverse operations and increased leverage operations have also appeared, and the overall risk to control is more casual, and the strategy has a certain degree of risk.

5. Conclusion

From the model construction and test results in this article, it can be seen that: first, the principal component analysis method is a more effective way to solve the multicollinearity problem, and it can use the data as much as possible while minimizing useless information; Second, distinguishing the inside and outside of the sample during the test can effectively avoid the use of future data, and the parameter values obtained from the test are also tested in the test outside the sample, which makes the validity of the model better verified. Third, multi-factor risk model has high practical application value, and its weight ratio ability can achieve higher returns while controlling risks as much as possible. Through the simulation of historical data, the model strategy has achieved an annualized return rate of 46% outside the sample. The energy sector defeated the CSI 300 Index with a lower valuation, which verified the application value of the multi-factor risk model and provided a reference for the further development of the multi-factor model and investors' asset allocation.

References

[1] Jia Xiujuan. Support vector machine quantitative stock selection based on random forest[J]. Regional financial research, 2019 (01): 27-30.

[2] Hu shuwen, Xu Jianwu. Application of principal component analysis and factor analysis in Chinese stock evaluation system [J]. Journal of chongqing university of technology (natural

science), 2017, 31(05): 192-202.

[3] Liu Zhaode, Zhan Qiuquan, Tian Guoliang. A review of comprehensive evaluation of factor analysis [J/OL]. Statistics and Decision, 2019 (19): 68-73 [2019-09-30]. https: //doi.org/10.13546 /j.cnki.tjyjc. 2019.19.015.

[4] Liao Li, Zhao Feng, Li Gefeng. Portfolio risk control method and risk factor identification based on multi-index model [J]. World Economy, 2003 (09): 44-49.

[5] Zhang Weinan, Lu Tongyu, Sun Jianming. Prediction and optimization of support vector machine in multi-factor stock selection [J]. Application of Electronic Technology, 2019,45 (09): 22-27.

[6] Zhao Shengmin, Liu Xiaotian. Introducing a multi-factor model of investor preference: an analysis based on the perspective of prospect theory [J]. China Economic Issues, 2019 (02): 106-121.

[7] Zhou Liang. Research on multi-factor stock selection strategy based on quantile regression[J]. Journal of Southwest University (Natural Science Edition), 2019,41 (01): 89-96.