Journal of Financial Risk Management, 2020, 9, 23-43
https://www.scirp.org/journal/jfrm
ISSN Online: 2167-9541
ISSN Print: 2167-9533

# Quantitative Stock Selection Strategies Based on Kernel Principal Component Analysis

**Meiyi Zhou*, Lianqian Yin**

Department of Finance, International Business School, Jinan University, Zhuhai, China
Email: *zhoumy321@126.com

**How to cite this paper:** Zhou, M. Y., & Yin, L. Q. (2020). Quantitative Stock Selection Strategies Based on Kernel Principal Component Analysis. *Journal of Financial Risk Management, 9,* 23-43.
https://doi.org/10.4236/jfrm.2020.91002

**Received:** January 11, 2020
**Accepted:** March 7, 2020
**Published:** March 10, 2020

## Abstract

Any stock is exposed to many different risk factors simultaneously. Collinearity among risk factors often makes it difficult to identify effective factors. Based on multi-factor quantitative stocks selection, 60 factors are selected including the fundamental, technical, macroeconomic factors and so forth. Then we map, process, and identify the data of Shanghai and Shenzhen 300 constituent stocks in high-dimensional space and extract characteristics of factors by kernel regression. The required model factors are determined by the variance contribution rate and the sample data are transformed by the feature vectors and kernel function, which are regressed with the stock returns to construct a multi-factor stocks selection model. The back test of historical data in 2017 indicates that the model has a lower back test and a higher return than Shanghai and Shenzhen 300 Index. Based on the bootstrap method, the robustness of the model is verified in the robustness test finally. The result shows that the combination of kernel principal component analysis and multi-factor stocks selection model is well verified, which can beat the market with high probability and effectively overcome the problem of random stocks selection.

## Keywords

Multi-Factor Stocks Selection, Bootstrap Method, High-Dimensional Mapping, Kernel Function

## 1. Introduction

The development of quantitative investment funds in China stays in the initial and developing stage. As the difficulty of fund performance beating the market is gradually rising, the major brokerage firms and investment funds began to invest in the research of quantitative investment. According to statistical data from

Wind, there are 63 quantitative fund products existing in China as of March 31, 2018, of which there are 61 funds which can outperform index returns in the same period. Quantitative funds in China will usher in the development boom. Stock index futures and margin trading business are launched in 2010 in China, providing investors with short tools, which can be used by investors to avoid systemic risks and obtain excess returns at advantage of the hedging mechanism. With the continuous improvement of securities market trading mechanism in China, quantitative investment will be developed rapidly. Many quantitative investment strategies have been verified in overseas markets. With the perfection of the securities market trading mechanism and the introduction of financial instruments, quantitative investment in China will have a big space for development. This paper intends to construct a multi-factor stocks selection model by extracting principal components by kernel function, and to provide a new idea for constructing multi-factor stocks selection model by quickly and accurately processing a large number of high-dimensional nonlinear classification problems, and also to provide with a good reference to method selections for other investors.

The research objective is that the paper tries to achieve the goal of processing large amounts of high dimensional nonlinear classification problems fast and accurately by adopting a new method to build a multi-factor stocks selection model. Additionally, this paper intends to provide a new thought for the research field of multi-factor stocks selection model and provides a new method for investors to choose good-performance stocks. Furthermore, based on previous studies of this kind, this paper attempts to improve the accuracy and profitability of the stocks selection model by fully considering a wide range of factors including environmental, micro, subjective and objective ones that affect stock prices.

The first part of this paper is the introduction part, which mainly introduces the background of the research and illustrates the research objectives of this paper. The second part is to review the related literature both at home and abroad. The third part mainly introduces the research design, which includes the whole research idea, the kernel method and stocks selection model. The highlight of the research is dimensionality reduction of factors by the kernel function the theoretical basis, including the development and theoretical foundation of the multi-factor model at various stages. Simultaneously, it introduces the principle, advantages, disadvantages and basic expressions of the kernel function, and explains why kernel function is applied to this paper. The fourth part is the core of this paper, which mainly introduces the process of building up a multi-factor quantitative model, including data collection and sorting, selection in factor library, standardization of each indicator, dimensionality reduction of factors by the kernel function, the establishment of a regression equation, selection of stock portfolios and analysis of the model results. The fifth part is to carry out the back test and non-randomity test of the model examining the feasibility of the model in practical application. The last part is a summary of the paper, explaining the advantages of the multi-factor stocks selection model in practical application and proposing the improvement direction of this paper.

## 2. Related Literature Review

Foreign research on multi-factor quantitative investment models is much earlier. The first basic research focused on the fundamental factors of companies. Sharpe (1964) proposed the CAPM model which finds that capital gains are not only affected by the risk-free investment income, but also by the market risk premium. Ross (1976) proposed arbitrage pricing theory and extended it (the CAPM theory) from the univariate model to the multivariate model. Asness (1997) first thought that not only fundamental data but also technical indicators have an impact on stock prices. After considering the impact of fundamental and technical factors on stock returns, the introduction of industry factors into multi-factor quantitative models has also become a trend. In this aspect, Albadvi & Chaharsooghi (2013) compared the impacts of different industries on excess returns by studying the German stock market and drew a conclusion that the category of industries is also one of the significant factors affecting stock returns.

Many domestic scholars have also conducted a lot of researches on multi-factor stocks selection strategies. Lin (2004) believed that China's securities market can also use fundamentals to conduct multi-factor stocks selection and conduct empirical analysis. Wang (2005a) divided the influencing factors of income into three different categories for analysis and considered that market size is the biggest factor affecting market returns. Wang (2005b) comprehensively used value factors and technical factors to construct stock portfolios for empirical analysis. Yuan (2008) perfected Wang Cheng's modeling method by combining technical and fundamental data. According to certain rules, five unconventional indicators were obtained, and the corresponding evaluation system was determined. Lin, Dai, & Ge et al. (2011) conducted a static analysis of the CSI 800 constituent stocks and selected 9 remarkable and significant style factors. Style factors and industry factors taken by the Shenwan first-class industry classification are used to construct multiple linear regression equation, where the retracement results show that the annualized excess return is about 10.64%. Liu (2012) examined the investment ability of stocks from the four dimensions of growth, valuation, financial quality and momentum, and selected eight indicators such as growth rate, operating cash flow growth rate, and return on equity (ROE) and so on to construct a factor selection model by testing the validity and stability of 25 factors, which certified good performance of the model in various aspects of the market. Ding (2012) systematically developed a set of basic procedures for quantitative stocks selection, which included the factors that have been confirmed by the market to affect stock returns into the self-built factor pool, and selected 30 factors to analyze empirically for further filtering effective factors.

With the development of neural networks and data mining in recent years, multi-factor based quantitative stocks selection models have been further developed. Based on the kernel principal component-genetic algorithm, Su & Fu (2013) proposed an improved artificial intelligence stocks selection model of supporting vector machine (KPCA-GA-SVM) and extracted the stocks of the Shanghai and Shenzhen stock markets for empirical analysis of the stock selection and prediction accuracy of the model in the short term and medium-term. Liu, Xia, Hu, & Lin (2016) constructed a GARP stocks selection model through the comprehensive scoring method dynamically predicted stock prices in order to generate timing signals with the method of the Markov chain. Wang, Cao, & Chen (2016) analyzed the index system adopted by the domestic and international quantitative stocks selection models, and used the index correlation analysis method to propose an eight-factor stocks selection model index system. Based on the sample data of 200 stocks in March 2013, a more accurate forecast for stocks in April 2013 was achieved with the method of the random forest algorithm. Based on the BARRA multi-factor model, Shi (2017) obtained different weights of each factor with the method of the random forest to avoid equal weight synthesis and result in information distortion, and extracted the information in the same factors with PCA, which avoids the overlap of information among style factors to cause estimation bias. The results showed that the system improves the accuracy and efficiency of model testing and application.

# 3. Research Design

## 3.1. Research Process

To begin with, the research method used in this paper is briefly described as follows: with the purpose of establishing a model that reflects the relationship between stock returns and various factors, we first need to consider which stocks to be chosen for analysis. Secondly, we should find out the relevant factors that affect stock returns. Then we determine the relationship between stock returns and related factors. Finally, relevant data are selected to test the feasibility of the model. In the aspect of stock pool selection: This paper takes all the constituent stocks of Shanghai and Shenzhen 300 Index in 2016 as the benchmark for the stock pool, statistically analyzes the data from 2010 to 2016, excludes stocks that have been missing data for more than 10 months, and supplement the missing data of the remaining stocks by Lagrangian Interpolation.

In the aspect of establishing the factor pool: This paper adopts the tree-like method, that is, firstly classify the influencing factors, and then classify the sub-categories into sub-categories. The factors in this paper are classified into five categories: fundamentals, technical indicators, macro, investor sentiment and analyst prediction. The secondary classification of factors, such as the fundamentals, includes scale factors, profit factors, capital structure factors, risk factors, etc. Then, the most representative indicators are included in the factor pool, where the secondary classification of factors such as scale factor and profit factor can be further divided into three levels. In the regression analysis stage: according to the traditional idea, the redundancy factor is directly eliminated due to the high autocorrelation among the secondary classification factors, and the most influential indicator in the second classification of factors is selected as the regression factor. This paper takes a different approach where the redundancy factors are not directly eliminated. Instead, it uses the kernel function to map the factors to the high-dimensional space, performs nonlinear regression to extract the principal components, and selects the principal components as the regression factors according to the contribution rate of the feature root cumulative variance, so as to maximize the use of the information contained in each indicator. Based on the dimensionality reduction data, the stock returns and factors are regressed, and the validity of the model is calculated by a statistical test. Then, the stock returns are predicted and the returns of each stock in the stock pool are ranked.

According to the study of the optimal number of holding shares launched by Li Wendi, the stocks ranked in the top 50 should be held for investment. In the model retracement stage: based on the sample data in 2017, the multi-factor quantitative model is back tested and compared with the Shanghai and Shenzhen 300 stock index during the same period to calculate the probability that the model can outperform the market. Then, randomly select 50 stocks to construct a portfolio with the method of bootstrap and calculate the combined returns to test the technical and non-random nature of the model, and finally draw the conclusions.

## 3.2. The Theoretical Basis of Multi-Factor Stocks Selection Model

Model (CAPM) proposed by William Sharpe in 1964. This model mainly introduces the relationship between stock returns and risk, which means that the return rate of stocks is equal to the risk-free rate plus risk compensation determined by the coefficient of $\beta$:

$$E(r_i) = r_f + \beta_{i,m}\left[E(r_m) - r_f\right] \tag{1}$$

$$\beta_{i,m} = \frac{Cov(r_i, r_m)}{Var(r_m)} = \frac{\sigma_{i,m}}{\sigma_m^2} \tag{2}$$

Among them, $\beta_{i,m}$ indicates the sensitivity of stock returns to market risk changes, and measures the systemic risk of stocks. The capital asset pricing model illustrates that the risk of stock is only related to market risk.

Since the Capital Asset Pricing Model assumes that the market is a fully efficient market, which is too strict, Ross extended it in 1976 and proposed the arbitrage pricing theory (APT). The theoretical basis of the model is no arbitrage equilibrium which relaxes the assumptions about the market. It believes that the market does not need to have an effective portfolio and the stock's return comes up with a series of linear combinations of related factors. The expression of the model is as follows:

$$r_{it} = r_f + b_{j1}RP_1 + b_{j2}RP_2 + b_{j3}RP_3 + \cdots + b_{jn}RP_n \tag{3}$$

Among them, $r_f$ refers to risk-free income, $RP_j$ refers to the *j*-th factor affecting stock returns.

Although the arbitrage pricing model has relaxed the requirements for the market, it cannot actually determine the factors affecting the stock returns. Therefore, Fama-French proposed a three-factor model and specialized the factors in 1993. They believed that the excess returns of stock of listed companies are determined by the risk exposure in three factors, such as market portfolio, company size and market value ratio, whose form is as follows:

$$r_{it} = r_f + \beta_{it}\left(r_m - r_f\right) + s_{it}\left(SMB_t\right) + h_{it}\left(HML_t\right) + \varepsilon_t \tag{4}$$

Among them, $r_f$ represents risk-free income. $\left(r_m - r_f\right)$ represents excess return in the market. $SMB_t$ represents excess return brought by the company's scale. $HML_t$ represents excess return brought by the book value ratio. $s_{it}$ and $h_{it}$ represent the risk exposure of the two factors at the moment of *t*-th.

Under normal circumstances, the smaller the company size, the higher the company's stock returns. There is a negative correlation between the two. The higher the book market value ratio, the higher the company's stock returns. There is a positive correlation between the two. This is because the smaller the scale, the higher the book value ratio, the greater the company's risk, so the higher the profit.

At present, the multi-factor stocks selection model is fully expanded on the basis of the three-factor model. A series of factors are selected as stocks selection criteria and can be expressed as:

$$r_{it} = r_f + \sum_{j=1}^{k}\left(r_{pjt} - r_f\right) * \beta_{ijt} + \varepsilon_t \tag{5}$$

Among them, $r_{it}$ represents the rate of return of the *i*-th stock at the moment of *t*-th. $\beta_{ijt}$ indicates the sensitivity of the *i*-th stock to the *j*-th factor at the moment of *t*-th. $\varepsilon_t$ is the error of the *i*-th stock at the moment of *t*-th. This paper is based on this model to improve the multi-factor stocks selection model.

### 3.3. Kernel Method

Kernel methods are a kind of pattern recognition algorithm, which is mainly applied to learn and find out the mutual relationship in a set of data. It has an obvious effect on identifying the nonlinearity among data. The principle is that separable linear variables are mapped to high-dimensional feature space by some forms of function.

The price of the stock is affected by many cross factors, and there is a commonality among nonlinear data. This paper transforms each influencing index and extracts the main components with the method of the kernel function.

The core of the kernel function lies in the construction of the inner product among the data. A suitable inner product is a key to solving a practical problem. Many excellent scholars have studied the kernel function and found some commonly used kernel functions as follows: Linear Kernel.

$$k\left(x_i, x_j\right) = x_i * x_j + c \tag{6}$$

1) Gaussian Kernel

$$k\left(x_i, x_j\right) = \mathrm{e}^{-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}} \tag{7}$$

2) Polynomial Kernel

$$k\left(x_i, x_j\right) = \left[a\left(x_i * x_j + c\right)^d\right] \tag{8}$$

3) Sigmoid Kernel

$$k\left(x_i, x_j\right) = \tanh\left[a\left(x_i * x_j + c\right)\right] \tag{9}$$

4) ANOVA kernel function

$$k\left(x_i, x_j\right) = \sum_{k=1}^{n} \mathrm{e}^{\left[-\sigma\left(x_i^k - x_j^k\right)^2\right]^d} \tag{10}$$

The idea of kernel function first appeared in the field of machine learning. After the kernel function is introduced, since the support vector machine does not need to know the specific mapping form of data, it can process nonlinear data in high-dimensional space quickly and efficiently, thus effectively reducing the workload of the computer. In addition, the kernel function is also favored by the field of artificial intelligence and machine learning. This is because the kernel function can accelerate the calculation of the algorithm and find the optimal solution faster. The kernel function can be divided into two parts: kernel function construction and algorithm design, which is a modular algorithm. Therefore, it can be combined with different algorithms to popularize and broaden the scope of application of kernel function in various fields. By combining the kernel function with the traditional algorithm, not only the speed and practicability of the traditional algorithm can be further improved, but also the common linear algorithm is generalized to the high-dimensional nonlinearity one. This paper combines the kernel function, linear regression and principal component analysis to develop methods called kernel principal component analysis and kernel principal component regression.

## 4. Data Processing and Factor Selection

### 4.1. Data Collection and Candidate Factor Selection

In order to ensure the integrity and the authority of the data, the data used in this paper comes from the database of CSMAR. In this paper, the method of tree-like classification is used to capture the factors hierarchically. The core basis is to consider whether the factors affect the stock returns. In order to make the selected factors completer and more comprehensive, the candidate factors are classified into fundamental, technical, macro, predictive, emotional categories. The factors shown in Table 1 are a total of 60-factor indicators in14 factor categories that contain all aspects of the company. In this paper, the time window is set from January 2010 to December 2016. The stock pool is constructed by the Shanghai and Shenzhen 300 constituent stocks. The selection of the target is based on the following two reasons: Firstly, the market value of all stocks in the Shanghai and Shenzhen 300 index accounts for 60% of the total market value, which can represent the overall operation of the securities market in China. Secondly, the investment income of the model back test can be compared with the return rate of Shanghai and Shenzhen 300 index that represents the overall performance of the market, which enables us to directly and objectively compare investment results from the model. In the collection of data corresponding to the stock pool, this paper comprehensively considers various factors affecting stock returns and collects data on five aspects of the company, such as financial data of fundamentals, technical indicators, macro data, investor behavior, analyst forecasts and so on. The data come from the company's quarterly financial statements from 2010 to 2016, weekly stock market data, reports of industry, reports from securities analysts and so on. In terms of data with missing values, if the null value does not affect the overall analysis, it should be supplemented with the

Table 1. Atotal of 60 factor indicators for 14 factor classifications.

| Scale Factor | Valuation Factor | Growth Factor | Risk Factor | Macroeconomic Factor |
|---|---|---|---|---|
| The total market capitalization<br>Total number of shares<br>Number of outstanding shares<br>Owners' equity | Price Earnings Ratio (PE)<br>Price to Book Ratio (PB)<br>Market Value Ratio (PCF)<br>Market Sales Ratio (PS)<br>Net Assets Per Share (BPS) | Year-over-year growth rate of earnings per share<br>Year-on-year growth rate of total profit<br>Year-on-year growth rate of operating profit<br>Year-on-year growth rate of net asset income<br>Year-on-year growth rate of operating revenue<br>Year-on-year growth rate of net assets<br>Growth rate of net asset per share | Tobin Q value<br>Book-to-market ratio<br>Alpha, Beta,<br>Sharpe ratio<br>Traynor index<br>Jensen index<br>$R^2$ | GDP Index<br>PMI Index<br>1 year time deposit interest rate |
| **Operating Capacity Factor** | **Income Quality Factor** | **Index per share** | **Technical Factors** | **Emotional Factors** |
| Operating cycle<br>Total asset turnover<br>Inventory turnover | Common stock yield<br>Net profit/total profit<br>Total operating cost rate | Earnings per share (EPS)<br>Operating income per share<br>EBITDA Per share<br>Pretax profit per share | Turnover, Turnover rate<br>MACD, DMI, RSI,<br>ROC change rate<br>Random index (KDJ),<br>energy tide (OBV)<br>Standard deviation of STD, | PSY mood indicator<br>Sentiment indicator (ARBR)<br>Investor confidence index |

method of Lagrangian Interpolation. If the data is missing for more than 10 months, the data of the stock is expected to be removed from the stock pool.

The factors in this paper mainly include 60-factor indicators of 14-factor classes. See Table 1 for specific factor selection. Among them, the fundamental factors include: 1) Scale Factor: it mainly reflects the value of the company. According to the existing data, since the smaller companies have greater operational risks, there will be higher risk compensation. Thus, the company size is negatively correlated with the stock return rate. 2) Valuation Factor: it mainly represents the expectation of investors on the value of stocks. If investors have low expectations for stock while its profitability is conversely very strong, then this stock is often undervalued, which is likely to bring benefits to investors. 3) Repayment Factor: it judges the ability of enterprises to repay short-term debt and long-term debt, reflects the company's credit rating and business strategy from the side, which is the basis for evaluating the risk of the company. 4) Growth Factor: it judges whether the company has long-term development potential. If a company can always benefit investors, this company is bound to develop in the long run. 5) Capital Structure Factor: it mainly reflects the operating structure and the direct balance strategy between the risk and liquidity of the company. 6) Profitability Factor: it reflects the ability of the company to obtain income, which has an important influence on whether investors should buy this stock. 7) Operational Capability Factor: it measures the ability of the company to use management resources to obtain income. 8) Yield Quality Factor: High-quality stocks tend to have stable returns in the long term. 9) Risk Factors: it mainly focuses on the risk exposure of the company. 10) Per Share Indicators: it represents the potential of stocks. Ten secondary classifications of indicators above can be further divided into three levels.

The technical factors are the data generated by the flow of listed company stocks among investors in the secondary market, which is the most important factor that integrates historical price information with a great influence on the timing of quantitative investment. The technical factor mainly reflects the trading indicators of stocks, such as the relatively intuitive opening price, closing price, volume and so on. In addition, there are also some data-derived indicators such as the daily average line, the smooth moving average, the exponential smoothing average, and some indicators obtained through complex calculations, such as the Bollinger line, relative strength indicators and so forth. Macroeconomic factors are the descriptions of relevant policies and developments in an industry or the entire securities market or country, which can have a great impact on the stock price changes. At present, there is still information asymmetry in the introduction of relevant domestic policies. Therefore, this paper only considers the factors that can be obtained in a timely manner by the public investors, such as gross domestic product (GDP) and purchasing managers' index (PMI). The analyst predictors are the forecast of development trends in the securities market by the securities analysts from major domestic securities firms, which are mainly obtained by analyzing the company's investment strategy and the market status and combining macroeconomic development with various factors. Since the brokers have the most comprehensive and timely information and the most professional knowledge in the industry, this paper also incorporates the analyst predictor factor into the stocks selection model.

Behavioral Finance believes that the assumptions of traditional asset pricing models are difficult to satisfy. Investors are no longer rational individuals, but emotionally driven groups, which is supported by group effects, overconfidence effects and so on. Especially in recent years, scholars are conducting a further study of emotion research in investors, thinking that emotion is an important factor affecting stock price changes. At present, the research on emotional indicators is mainly based on the PCA method or market survey. This paper selects a psychological line PSY emotional index and ARBR popular willing index for quantitative analysis.

### 4.2. Stock Selection Model Based on Kernel Principal Component Regression

In this paper, 60 influencing factors are selected to construct the stocks selection model, but the factors under the same classification system have a high degree of linear correlation. In terms of the scale factor, the total share capital includes tradable shares and restricted shares, so the total share capital and circulating equity must inevitably overlap in the aspect of affecting stock prices. Direct regression will lead to collinearity, making the coefficients of each factor ineffective and causing the model to failure in prediction.

As for the solution to this problem, the traditional approach is to retain only the most influential factors in the same category, and directly reject the residual factors. However, it causes low utilization of the factor pool and can’t include all factors that affect the stock returns in this approach.

The idea of this paper is to reduce the dimensionality of all factors to extract the principal components, and to analyze the principal components and stock returns by regression analysis. The kernel function has the ability to map the data to a high-dimensional space and effectively process the nonlinear component among data. It is the highlight that the extracted principal components and returns of stock are analyzed by regression analysis making use of the high-dimensional space constructed by kernel function, which is called kernel principal component regression. This method can summarize all the factors affecting stock returns more comprehensively and overcome the problem of inter-factor collinearity.

The empirical analysis in this step is also the most important part of this paper.

When the data is centralized, the data satisfy the assumption of the centralization condition, that is:

$$\sum_{i=1}^{n}\Phi\left(x_i\right)=0 \tag{11}$$

Then, the covariance of data after the procession can be expressed as:

$$C^{\Gamma}=\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\Phi(x_i)^{\mathrm{T}} \tag{12}$$

The following formula can be obtained by derivation:

$$\lambda v=C^{\Gamma}v \tag{13}$$

Among them, $\lambda>0$ represents matrix eigenvalue, $v\in\Phi(\cdot)$ represents the eigenvector of the matrix. Multiplying both sides of the equation simultaneously with the kernel sample $\Phi(x_k)$, the equation is transformed into:

$$\lambda\left[\Phi(x_k)v\right]=\Phi(x_k)C^{\Gamma}v \tag{14}$$

Since the fact that $v\in span\left[\Phi(x_1),\Phi(x_2),\cdots,\Phi(x_n)\right]$, there must be an array like $a_i\,(i=1,2,\cdots,n)$ to make:

$$v=\sum_{i=1}^{n}\alpha_i\Phi(x_i) \tag{15}$$

Then,

$$\lambda\sum_{i=1}^{n}\alpha_i\left[\Phi(x_k),\Phi(x_i)\right]=\frac{1}{n}\sum_{i=1}^{n}a_i\left[\Phi(x_k),\sum_{j=1}^{n}\Phi(x_j)\right]\left[\Phi(x_j),\Phi(x_i)\right] \tag{16}$$

According to the definition of the kernel function, the data samples $x_i,x_j$ input in the high-dimensional feature space are determined by their inner product $\left[\Phi(x_j),\Phi(x_i)\right]$ with a kernel function. Therefore, the inner product matrix $[K]_{ij}$ is defined with $n$ as the sample capacity. The Formulas (12)-(13) can be expressed by kernel matrix as:

$$n\lambda\alpha=K\alpha \tag{17}$$

It can be clearly seen from the equation above, $\alpha=\left[a_1,a_2,\cdots,a_n\right]^{\mathrm{T}}$ represents the matrix eigenvector. Making $\lambda^*=n\lambda$, $\lambda^*=\left[\lambda_1^*,\lambda_2^*,\cdots,\lambda_n^*\right]$ represents the corresponding eigenvalue, which can derive that the *k*-th principal component in high-dimensional space represents the projection of the mapping $\Phi(x_i)$ of the sample points $x_i$ on the eigenvector $v_p$. The formula is expressed as:

$$\beta(x)_k=\left[v_k,\Phi(x)\right]=\sum_{i=1}^{n}\alpha_i^k\left[\Phi(x_i),\Phi(x)\right]=\sum_{i=1}^{n}\alpha_i^kK(x_i,x) \tag{18}$$

Among them, $k=1,2,\cdots,p$, where *p* represents the number of the principal component in the function construction.

The application of kernel principal component analysis to process data avoids the problem of excessive computation and a long time of analysis caused by the simultaneous inclusion of indicators in the traditional quantization process. At the same time, the data is processed in high-dimensional space to extract nonlinear features, which makes all of the data effectively utilized. Investment decisions can be made more clearly and accurately using this method to analyze stocks.

After extracting the principal components from the factor library, the components are ranked according to the variance contribution rate, and principal components from 1-st to *p*-th and the income is selected for Least Squares Regression Analysis. The model is expressed as:

$$y = \sum_{k=1}^{p} \omega_k \beta(x)_k + \varepsilon = \sum_{k=1}^{p} \omega_k \sum_{i=1}^{n} a_i^k K(x_i, x) + \varepsilon \tag{19}$$

Also, the model can be expressed as:

$$y = B\omega + e \tag{20}$$

Among them, $B$ represents matrix conversed by original data and is the mapping of original data in high-dimensional space eigenvectors, whose orthogonal relationship can be expressed as:

$$B = \Phi V = \sum_{i=1}^{n} a_i^k K(x_i, x), (k = 1, 2, \cdots, p) \tag{21}$$

The fact that $\Phi = [\Phi(x_1), \Phi(x_2), \Phi(x_3), \cdots, \Phi(x_n)]$ represents the raw data array transformed by kernel function. $V$ represents the previous *p*-th eigenvector matrix selected for the model. $y$ represents the rate of return. $\omega$ represents the matrix of regression coefficient. $e$ represents the matrix of residual. According to the principle of minimum variance, the estimated expression of regression coefficient can be expressed as:

$$\hat{\omega} = \left(B^{\mathrm{T}} B\right)^{-1} B^{\mathrm{T}} y = \Lambda^{-1} B^{\mathrm{T}} y \tag{22}$$

The basis of constructing the kernel function is to select the right kernel function. The commonly used kernel functions include polynomial kernel function (POLY), Gaussian radial basis kernel function (RBF), multi-layer perceptron (TANH) kernel function and ANOVA kernel function. In this paper, the above kernel functions are used to extract principal components of the data, and the kernel functions used in the stocks selection model are determined by analyzing the effects of different kernel functions.

The first stage is data preprocessing. Since the data has different units, the analysis of all the data in the same model will affect the weight of different indicators, making the results lack validity. Therefore, firstly, the dimension of 60 indexes collected data is reduced with the method of Z Standardization method to make the data obey a standard normal distribution with a mean of 0 and a variance of 1. The specific processing method is as follows:

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \tag{23}$$

The second stage is determining kernel function parameters. Map the normalized data to high-dimensional space using kernel functions, that is

$x_i \to \Phi(x_i) \in \Gamma$. Then based on the method of cross-validation selection, the polynomial kernel function parameter is determined to be 4, and the Gaussian radial basis kernel function parameter is 0.03.

The third stage is the inner product matrix calculation. The parameters determined in the second stage are substituted into the Matrix Function in Kernlab, and the kernel inner product matrix is calculated. The dimension of the matrix is *n*, and the number of samples is *n*. The expression is as follow:

$$K_{ij} = \left[\Phi\left(x_i\right),\Phi\left(x_j\right)\right] = \left[k\left(x_i,x_j\right)\right] \tag{24}$$

The fourth stage is the centralization of the inner product matrix. It can be known from the formula above that the matrix needs to be centralized by making its mean equal to 0 in order to obtain the eigenvalues and eigenvectors of $\Phi\left(x_i\right)$. The specific steps are as follows:

$$\bar{K} = K - I_nK - KI_n + I_nKI_n \tag{25}$$

The fourth stage is the eigen decomposition of the centralization matrix. the specific steps are as follows. According to

$$n\lambda\alpha = \lambda^*\alpha = \hat{K}\alpha \tag{26}$$

the eigenvalue $\lambda^*$ and eigenvector $\partial$ of the Kernel inner product matrix $\hat{K}$ are obtained. Make the length of the eigenvector $v_p$ in the high-dimensional space $\Gamma$ reflected by the function equal to 1, which is called Unit Orthogonalization. That is:

$$\begin{aligned} v_p * v_p &= \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ip}\alpha_{jp}\Phi\left(x_i\right)\Phi\left(x_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ip}\alpha_{jp}K_{ij} \\ &= \alpha_p K_{ij}\alpha_p = \lambda_p^*\left(a_p * a_p\right) = 1 \end{aligned} \tag{27}$$

Thus, $\left(\alpha_p * \alpha_p\right) = 1/\lambda_p^*$ is obtained.

The sixth stage is the determination of kernel function in the model. The obtained eigenvalues are arranged in descending order according to the variance size, and the variance contribution rates accumulated by each eigenvalue are calculated. The eigenvectors with more than 90% cumulative contribution rate of variance are used as the principal components extracted by the kernel function and the factors of the regression model. The cumulative variance contribution rates of the four kernel functions are shown in Table 2.

In Table 2, four kernel functions such as RBF, POLY, neural network ANOVA are used to extract the principal components. The first six principal components are extracted in total. The upper row of each function in the table refers to the corresponding characteristic value of the component, and the nether row refers to the cumulative variance contribution. According to the results from Gaussian kernel function extraction, the cumulative contribution rate of the variance corresponding to the sixth eigenvalue only reaches 31.6%, which means the extraction effect is so poor. As for the results from the polynomial kernel function extraction, the cumulative contribution rate of variance corresponding to the fifth eigenvalue reaches 92.2%, which shows five principal components effectively extract the more than 90% of the commonality of the nonlinear factors. In terms of results of the neural network kernel function extraction, the extraction is inferior to that of polynomial kernel function owing to the horizontal comparison of the cumulative variance contribution rate corresponding to each eigenvalue. The cumulative contribution rate of the last eigenvector of the Anova kernel function is 47.2%, thus the kernel function extraction result is poor too. Therefore, the dimensionality reduction effect is the most obvious in the polynomial kernel function, with the cumulative variance contribution rate of the first five principal components reaching 92%, which can represent most of the factors affecting stocks. Combining with the parameters determined by orthogonal units, the polynomial kernel function parameters are finally determined to be 4.

The seventh stage is data collection. According to the kernel function used to construct the model determined in the sixth stage, take the cumulative variance contribution rate of 90% as the reference line, and determined the number of extracted principal components be 5. Then, adjust eigenvector and select $a_1, a_2, \cdots, a_5$ according to the fifth stage to calculate the projection of the eigenvector in the centralized kernel inner production matrix $B = \hat{K} \cdot \left(a_1, a_2, \cdots, a_5\right)$ as the data after the dimension reduction.

The eighth stage is the construction of the stocks selection model. Take the yield rate of the Shanghai and Shenzhen 300 constituent stocks in the lag period as the dependent variable of the regression model, which is used to perform the Least Principal Regression with the five principal components extracted from the corresponding constituent stocks, and perform regression analysis with R language. The expression can be expressed as follow:

$$\hat{\omega} = \left(B^{\mathrm{T}} B\right)^{-1} B^{\mathrm{T}} y = \Lambda^{-1} B^{\mathrm{T}} y \tag{28}$$

Therefore, predictor expression for dependent variable is as follow:

$$\hat{y} = B\hat{\omega} = B\Lambda^{-1} B^{\mathrm{T}} y \tag{29}$$

The results of regression are shown in Table 3.

In Table 3, three coefficients of the principal components are negative, indicating that a unit principal component change will reduce the corresponding stock returns. Two principal component coefficients are positive, indicating that a unit principal component change will increase the corresponding stock returns.

Table 2. Effects of extracting principal component by kernel functions.

| *Kernel Principal Component* | *Comp_1* | *Comp_2* | *Comp_3* | *Comp_4* | *Comp_5* | *Comp_6* |
|---|---|---|---|---|---|---|
| Gaussian Kernel Function RBF | 23.753 | 13.759 | 10.852 | 8.631 | 6.386 | 5.847 |
| | 10.4% | 17.1% | 22.5% | 26.3% | 29.1% | 31.6% |
| Polynomial Kernel Function | 2.31E$^{+11}$ | 1.09E$^{+11}$ | 7.11E$^{+10}$ | 1.28E$^{+10}$ | 1.08E$^{+10}$ | 1.05E$^{+10}$ |
| | 52.6% | 76.3% | 86.5% | 89.9% | 92.2% | 94.3% |
| Neural Network Kernel Function | 28.316 | 21.755 | 16.326 | 11.328 | 8.351 | 3.185 |
| | 30.6% | 45.1% | 58.3% | 70.6% | 79.2% | 86.3% |
| Anova kernel function | 1.64E$^{+04}$ | 1.18E$^{+04}$ | 9.63E$^{+03}$ | 8.19E$^{+03}$ | 7.03E$^{+03}$ | 6.22E$^{+03}$ |
| | 12.6% | 21.5% | 29.7% | 36.4% | 42.8% | 47.2% |

Table 3. Parameter estimation of kernel principal component regression.

| | *Coefficient* | *Standard Deviation* | *Value of T* | *Value of P* |
|---|---|---|---|---|
| c | $-1.53E^{-03}$ | $5.31E^{-03}$ | −0.294 | 0.63859 |
| Comp_1 | $-4.58E^{-05}$ | $4.82E^{-06}$ | −10.315 | $<2e^{-16}$ |
| Comp_2 | $-6.92E^{-05}$ | $6.93E^{-06}$ | −9.846 | $<2e^{-16}$ |
| Comp_3 | $-2.49E^{-05}$ | $8.05E^{-06}$ | −3.182 | 0.00113 |
| Comp_4 | $4.63E^{-05}$ | $9.31E^{-06}$ | 5.634 | $1.32E^{-07}$ |
| Comp_5 | $4.81E^{-05}$ | $1.04E^{-05}$ | 5.018 | $1.18E^{-06}$ |

As can be seen from the value of *P*, with the significance level of 1%, the five principal components have all passed the OLS regression test, which shows the statistical magnitude of the principal components multi-factor regression model based on the kernel function is effective, and the model can be used to predict stock returns.

This chapter constructs the stock pool and factor pool under the consideration of various factors, then uses the kernel function to map 60 indicators in high-dimensional space and extracts the principal components with the method of the Kernel Principal Component, which is followed by the stage that the return rate of stock is used as the dependent variable to construct the multi-factor stocks selection model by the Kernel Regression Analysis with extracted principal components. Finally, the saliency of the model was well verified. It is proved that the model has a good fitting effect by the calculation of the value of F, MSE, the value of P and other indicators in the equation.

## 5. Comparison Test and Robustness Analysis

In the previous section, factor selection and processing were carried out, and the Shanghai and Shenzhen 300 constituent stocks were used as a stock pool to construct a multi-factor stocks selection model suitable for all constituents of the Shanghai and Shenzhen 300. When the model was backtested, the stock benchmark remained unchanged. Since the constituents of the Shanghai and Shenzhen 300 are constantly updated, abnormal ST stocks are automatically removed. Therefore, all constituents of the Shanghai and Shenzhen 300 are directly used as the retracement target.

This section mainly examines whether the model revenue can beat the market benchmark in actual application. When using the model to select stocks and open positions, the portfolio revenue is compared to the Shanghai and Shenzhen 300 Index. In the back-testing process, the first step is to determine the investment scale. The more the selected targets, the more the non-systematic risk of individual stocks can be significantly reduced by decentralization, but the investment funds need to be expanded accordingly. Based on the general investment strategy of foreign funds currently, the optimum number of stock portfolios is about 60. The core of this paper is to establish a multi-factor stocks selection model, so there is no further discussion on the scale of investment.

In this paper, the multi-factor stock selection model is used to predict the returns of stocks, and the stock returns are arranged in descending order, where the top 50 stocks are selected to form a portfolio. The selected companies are listed in Schedule 1. According to the industry classification in the Dow Jones index, it can be seen that the industrial sectors that make up the investment portfolio account for a large proportion, especially resource-based industries such as China Petroleum, Sinopec, China Railway and soon. Other industries also include the aviation industry and the electrical industry. As can be seen from the stock portfolios, the industry factor is a significant stock-picking signal among the influencing factors of the model. Then, the stock returns are back-tested, and the statistics show the comparison between the return of the portfolio and that of the Shanghai and Shenzhen 300 index in the same period as Figure 1.

It can be seen from Figure 1 that the portfolio selected by the multi-factor model effectively captures the upward trend of the benchmark in the Shanghai and Shenzhen 300 index in the same period. In addition, the stocks selection strategy makes significant use of all factor information. As the time window scrolls, it can be seen that the positive return difference between portfolio income and the Shanghai and Shenzhen 300 Index during the same period is significantly enlarged in the six investment stages. At the end of the model retracement, the combined net value was 1.41, that is, the annual portfolio income reached 41%, well ahead of the nearly 20 percent rise in the index over the same period. The portfolio income can be significantly higher than that of the Shanghai and Shenzhen 300 Index during the same period, which fully demonstrates the multi-factor stocks selection model based on the kernel function has a good effect in the practice of A-share market in China. In addition, apart from the measurement of return in the evaluation of the model, the size of the portfolio

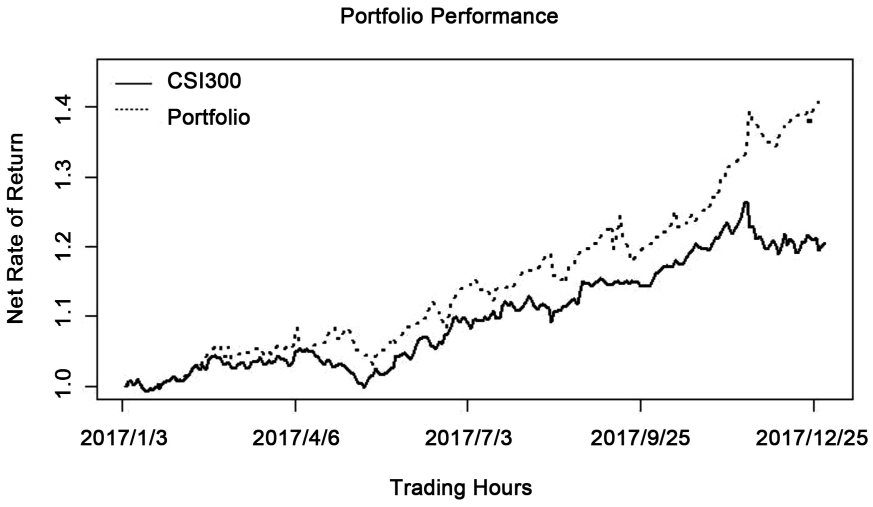

Figure 1. Comparison of portfolio income.

risk is also an important factor. A good portfolio can minimize the risk of investment on the basis of guaranteed returns. Three indicators such as the coefficient, Sharpe ratio, maximum retracement rate is used in the measurement of portfolio risk. The coefficient of $\beta$ refers to the sensitivity of investment income to market risk. The larger the value, the greater the impact of the market on the investment income. The Sharpe ratio represents the excess return that judges the risk of the combined unit. The larger the value, the more reward it can get in the risk. The maximum retracement rate shows the maximum retreat value of the return rate when the product net value reaches the lowest point at any historical time point in the selected period. The specific formula can be expressed as:

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p} \tag{30}$$

Among them, $R_p$ represents the combined annual yield. $R_f$ represents market risk-free rate of return. $\sigma_p$ represents the combined annual volatility.

$$\text{Max Drawdown} = \frac{\text{Max}\left(P_x - P_y\right)}{P_x} \tag{31}$$

Among them, $P_x$, $P_y$ represents the total value in the other combined day, with $y > x$.

After calculation, the combination and benchmark of risk indicators are shown in Table 4.

It can be seen from Table 4 that the value of $\beta$ in the constructed portfolio is 0.183, which is less than that of the market benchmark (1) in the same period, indicating the system risk of the portfolio is considerably less than the market average and the adverse effect is that the beta income caused by the system risk will be reduced. The Sharpe ratio of the portfolio is 3, which is far higher than that of the market (1.367), demonstrating the portfolio can achieve significant alpha gains beyond the bear market. The maximum retreatment rate of the portfolio is 0.045, which is marginally lower than the market level (0.061), showing the curve of portfolio income is smooth and portfolio income can effectively overcome market noise.

Although the back test empirically verifies the ability of portfolio income to outperform the market with a high probability, it does not indicate whether the yield result is derived from the technical or random nature of the model for the reason that random stock selection may still result that the return rate of combined portfolio is still higher than that of the market. Therefore, the randomness of the model is tested to illustrate the robustness of the model itself in this section. In order to prove the robustness of the model, the bootstrap method is used by multiple stochastic construction combinations to simulate the income results of stochastic combinations. Then the yield of a selected stock is compared to determine the distribution range of the stock yield rate and prove the non-randomness of the model. The specific steps are as follows: First of all, ran-

Table 4. Risk indicators of portfolio.

| | portfolio | market benchmark |
|---|---|---|
| The coefficient of $\beta$ | 0.183 | 1 |
| The Sharpe ratio | 2.974 | 1.367 |
| Maximum retracement | 0.045 | 0.061 |

domly determine the portfolio of 50 stocks based on the Shanghai and Shenzhen 300 Index and evenly distribute the funds to the constituent stocks in the portfolio. Then, perform the retracement process to calculate the income in each position respectively and then sum up equally to the total income of the combined portfolio, which is followed by the stage that 1000 random portfolios are constructed by Bootstrap Method. Next, the portfolios are constructed by randomly sampling the constituents in the stock pool. The combined income is determined by the retracement process. What's more, the portfolio can be constructed by combining with the sample function and circularly using the loop statement 1000 times to build up the portfolios in order to obtain the simulated data of portfolio yield rate. Eventually, a frequency histogram of return rate in a random portfolio is constructed. The yield of 1000 random portfolios calculated in the previous two steps is divided into 15 classifications, which can be described in a frequency distribution diagram as shown in Figure 2.

It can be seen from Figure 2 that the investment returns of the random portfolios closely obey the normal distribution. Based on the frequency of the simulated data points, the investment returns of the random portfolio are mainly concentrated between 1.02 and 1.14, accounting for nearly 50%. The highest range of investment yield is above 1.34, accounting for 0.003. Moreover, the minimum yield rate of the random portfolio reaches 0.8, accounting for 0.001. The final step is to determine the interval of yield rate in selected stocks, whose return at the end of the strategy retracement is 1.41. Compared with the histogram of yield rate in random portfolios, it can be seen that the yield of the portfolios selected by the model is in the highest range of the yield distribution, which fully illustrates the robustness of the model itself, eliminates the problem of random stocks selection, and proves the technicality of the model.

## 6. Conclusion

In this paper, the constituents of Shanghai and Shenzhen 300 Index are used as stock pools and the factor pool including 60 indicators is formed by comprehensive factors. Then the principal components are extracted based on the kernel function and the multi-factor investment stocks selection model is constructed. Finally, the model is back-tested and tested for randomness. In the whole process, the innovations and conclusions can be attributed to the following points:

1) In the aspect of selecting the factor library, this paper combines the research results of the previous scholars, and incorporates the fundamental factors

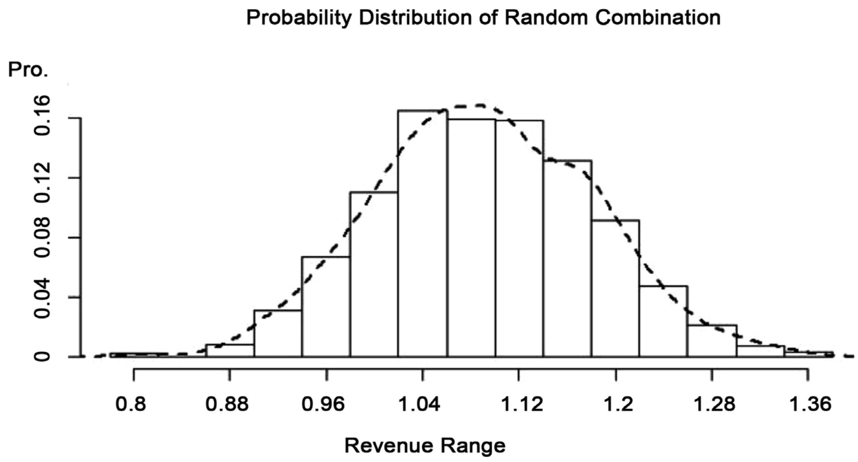

Figure 2. Frequency distribution histogram of random portfolios.

including growth ability, scale factor, and technical factors such as volume and turnover rate into the factor pool. At the same time, the investor sentiment indicators representing behavioral finance, the macroeconomic indicators referring to market factors, and the analyst forecast indicators indicating future development trends are included in the factor pool, which considers as many factors affecting the portfolio as possible.

2) The influencing factors of stocks trade are complicated. In this paper, the method of extracting principal components in high-dimensional space by kernel function to map raw data takes the place of the traditional method of directly eliminating redundant factors, which retains the comprehensiveness of information in candidate factors.

3) The dimension reduction by using the representative principal component and stock return rate for regression analysis is significantly more effective than that with the original data and stock return rate. The effect of parameters and the indicators of the model are still in line with the regression requirements. Although the built investment portfolio has a good effect on application, there are still some shortcomings as follows: the transaction cost of stocks is ignored. Also, multi-factor quantitative investment stock selection strategy is taken into account while trading timing is left out of the consideration. In practical application, funds or brokers will combine various investment strategies to obtain excess returns, or use stock index futures, share options or derivatives for risk hedge, which becomes the direction of improvement in this paper.

## Acknowledgements

First of all, we would like to express our gratitude to research on arbitrary fuzzy reconstruction of complex data and jumping detection method of National Natural Science Foundation of China (6507010083), the support for the risk contagion effect and the buffer mechanism in the economic cooperation between South Asia and China in the 12th Five-Year Plan for Philosophy and Social

Sciences in Guangdong province in China (GD14XYJ30), research on the Construction of English Course of Financial Engineering and International Education, and National University Students Extracurricular Academic Science and Technology Competition, which are of great help and importance in making the thesis a reality.

In addition, we owe a special debt of gratitude to all the professors who helped us during the writing of this thesis for offering us valuable suggestions in the academic studies.

Last but not least, our gratitude also extends to our family and friends who have been assisting, supporting and caring for us all of our lives.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

Albadvi & Chaharsooghi (2013). *Risk-Reward Views: Unlocking the Full Potential of Fundamental Analysis.* Morgan Stanley, Global Research.

Asness, C. S. (1997). The Interaction of Value and Momentum Strategies. *Financial Analysts Journal, 53,* 29-36. https://doi.org/10.2469/faj.v53.n2.2069

Ding, P. (2012). *Quantitative Investment #: # Strategy and Technology.* Shanghai, China: Publishing House of Electronics Industry.

Lin, D. Z. (2004). An Empirical Analysis of the Performance of Value Investment in China's Stock Market. *Finance and Economics, S1,* 271-274.

Lin, X. M., Dai, J., & Ge, X. Y. (2011). The Third Road: The Concept of Cordyceps Sinensis Investment Fund. *Capital Markets, 4,* 56-59.

Liu, Y. (2012). *Empirical Research on Factor Stock Selection Model in Chinese Market.* Shanghai, China: Fudan University.

Liu, Y., Xia, S. Y., Hu, S. R., & Lin, S. L. (2016). Research on GARP Quantitative Stock Selection and Markov Chain Timing Strategy. *Finance & Economy, 5,* 66-71.

Ross, S. A. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory, 13,* 341-360. https://doi.org/10.1016/0022-0531(76)90046-6

Sharpe, W. F. (1964). Risk-Aversion in the Stock Market: Some Empirical Evidence. *The Journal of Finance, 20,* 416-422. https://doi.org/10.1111/j.1540-6261.1965.tb02906.x

Shi, W. F. (2017). *Application of PCA and Random Forest in BARRA Quantitative Hedging Model.* Xi'an, China: Xi'an University of Science and Technology.

Su, Z., & Fu, X. Y. (2013). Improved Kernel Principal Component Genetic Algorithm and SVR Stock Selection Model. *Statistical Research, 30,* 54-62.

Wang, C. (2005b). Empirical Research on China's Stock Market Value Strategy. *World Economic Papers, 6,* 32-38.

Wang, S. Y., & Cao, Z. F., & Chen, M. X. (2016). Research on the Application of Random Forest in Quantitative Stock Selection. *Operations Research and Management, 25,* 163-168+177.

Wang, X. L. (2005a). *Multi-Factor Pricing Model Theory and Testing in China's Stock Market.* Wuhan, China: Wuhan University.

Yuan, J. (2008). *Stock Price Analysis.* Chengdu, China: Southwest Jiaotong University.

## Abbreviations

GDP: Gross Domestic Product
PMI: Project Management Institute
EBITDA: Earnings before Interest, Taxes, Depreciation and Amortization
MACD: Moving Average Convergence/Divergence
DMI: Directional Movement Index
RSI: Relative Strength Index
ROC: Receiver Operating Characteristic
PSY: Psychology