Implementing stock prediction using partial least squares

Chengfei Peng

Dalian University of Technology, Dalian, Liaoning, China

3206826217@qq.com

Abstract. This paper demonstrates that a considerable number of investors are entering the stock market as the economy displays signs of post-epidemic recovery. Nevertheless, even though the stock market acts as the foundation of the financial sector, not all stockholders can achieve significant profits from it. Yet, stock prices in the stock market have been challenging to keep rational and lack a stable pattern, marked by a notable level of uncertainty, which complicates investment decisions. Consequently, this paper aims to employ scientific mathematical techniques for the analysis and prediction of stock prices, with the goal of achieving a more comprehensive grasp of the stock market's general trend, thereby promoting prudent investment planning.

Keywords: stock forecasting; partial least squares (PLS) algorithm; investment; finance.

1. Introduction

Numerous prediction methods within mathematical modeling, including time series prediction theory and gray prediction theory, can forecast stocks [1]. Nonetheless, these methods exhibit various shortcomings. Bai Lu has described the shortcomings of time series models and gray forecasting models in her article [2]. For example, the time series forecasting model involves complex parameter selection, challenging parameter adjustments, and poor model interpretability. Similarly, the gray forecasting model exhibits inadequate medium and long-term forecasting accuracy, typically suitable only for exponential growth prediction [3]. Thus, this paper primarily employs the partial least squares algorithm for forecasting, aiming to mitigate multicollinearity among factors [4]. Through stock simulation experiments, we demonstrate that the partial least squares method can approximate actual values effectively, even when stock data is limited.

Partial Least Squares Regression (PLSR) is a regression modeling technique that extends the least squares approach for complex relationships between dependent variables and multiple independent variables [5]. It is able to address the interdependence between two sets of highly correlated variables, allowing the researcher to use one set of variables (independent or predictor variables) to predict another set of variables (dependent or response variables). When linear correlations between a large number of variables are involved and there are relatively few observations or the number of sample points is insufficient for the number of independent variables, PLSR can effectively deal with the problem of multicollinearity[6]. In addition, the PLSR model includes all the independent variables, which makes the interpretation of regression coefficients more intuitive, and its high predictive accuracy helps to provide a clearer qualitative interpretation.

2. Model Formulation

The essential independent variables include the P/E ratio, trading volume (shares), SSE index, KDJ indicator (K value), KDJ indicator (D value), mean, and standard deviation of stock data. The dependent variable is the closing price ($) [7]. Typically, this situation is addressed using the Ordinary Least Squares (OLS) method to establish seven regression equations involving seven independent variables. However, this approach can lead to the creation of numerous preconditions and assumptions, and there is no assurance that real-world data aligns with these assumptions. Therefore, Principal Component Analysis (PCA) can be employed to classify various data types, reducing the number of independent variables. However, PCA has its inaccuracies. The primary concern is its reliance on intuition, which can jeopardize the accuracy of categorization. Therefore, it is advisable

to consider the contribution of principal components to the dependent variable prior to categorizing the independent variables. This concept underlies the principle of Partial Least Squares (PLS) [8]. Based on the preceding explanation, it becomes apparent that Partial Least Squares (PLS) can be viewed as a fusion of Principal Component Analysis (PCA) and Regression Analysis (RA). Its most distinctive aspect is that the "Principal Component Analysis" within Partial Least Squares (PLS) considers the contribution to the dependent variable.

PLSR offers significant advantages in stock prediction modeling due to the multifactorial nature of the stock market and the frequent presence of multivariate covariance. Traditional regression models struggle with these challenges. PLSR, on the other hand, can simultaneously handle multiple interrelated factors, addressing the covariance issue by maximizing the covariance between independent and dependent variables, enhancing model fit. Furthermore, PLSR enables dimensionality reduction of independent variables during modeling while preserving crucial information. This simplifies the model and mitigates dimensionality-related issues. This enhances model computational efficiency, lowers the risk of overfitting, and aids in pinpointing the most influential features driving stock price fluctuations. Moreover, financial markets are characterized by noise and volatility, and PLSR models exhibit resilience to noise and outliers in the data. This enhances the model's suitability for handling high-uncertainty, volatile market data. Additionally, PLSR excels at accurately capturing the influence of diverse factors on stock prices, particularly when dealing with multiple correlated factors and nonlinear associations. This facilitates the development of more informed investment strategies for investors. Moreover, PLSR is a versatile modeling method not constrained by particular data distribution assumptions, rendering it applicable across diverse data types and market scenarios. PLSR generates components that elucidate how individual factors contribute to stock prices, enhancing model transparency and facilitating well-informed investment decisions.

3. Empirical Study

This study focuses on Shanghai Pudong Development Bank, referred to as "Pudong Development Bank," and collects stock trading data from January 2022 to October 2023 for a thorough analysis. Our study primarily examines the stock trading performance of Pudong Development Bank throughout this period. Through data analysis, we seek to comprehend the fluctuations in the bank's stock prices, trading volumes, and potential market trends. This analysis provides valuable insights into the bank's position and performance in the financial markets. A comprehensive analysis of Pudong Development Bank's stock data will be carried out using data processing and analysis tools such as SPSS and Matlab.

Data preprocessing begins with consolidating stock data from various files and sources into a unified dataset. This involves standardizing timestamps, identifiers, and data formats to facilitate effective data analysis. Immediately after data cleansing, we identify and address missing data, outliers, and errors. This process may include interpolation, deletion, or data filling to maintain data integrity. Because data sources may employ varied units of measurement and ranges, it becomes necessary to normalize the data to a common scale, facilitating comparisons and modeling. Data smoothing is a common step to reduce data volatility. This is achieved by applying a moving average to create smoother time series data. Feature engineering techniques are employed in preprocessing to extract valuable features or metrics from the raw data, enhancing the characterization of stock price changes and related factors.

The ideal quantity of principal components is initially ascertained through an evaluation of explained variance and VIP (cumulative predictive importance). This phase seeks to identify the appropriate number of principal components required for the efficient capture of data variability and information. Subsequently, the composition of principal components is extracted from the component matrix table [9]. After determining the quantity of principal components, we extract these components from the component matrix table to discern their weights and relationships among various variables,

thereby unveiling the underlying data structure and patterns. The table of factor loading coefficients is subsequently employed to evaluate the significance of individual variables. During this phase, we employ the factor loading coefficient table to evaluate the contribution and importance of each variable within the principal components. This process aids in identifying the variables pivotal to the formation of principal components, thereby enhancing our comprehension of the data's inherent characteristics [10]. This culminates in the ultimate standardized formula for Partial Least Squares Regression (PLSR). In conclusion, we create and present the ultimate standardized formula for PLSR to facilitate additional data analysis and predictive tasks. This formula equips us with a robust tool for modeling and prediction, leveraging the significance of principal components and variables identified in prior steps.

3.1. Analysis

The first principal components υ_1 and ν_1 of the independent and dependent variables are extracted, where X is a linear combination of the set of independent variables $X=[x_{1,}x_{2,\ldots\ldots}x_{m,}]^T$ linear combination, Y is the dependent variable set $Y=[y_{1,}y_{2,\ldots\ldots}y_{p,}]^T$ linear combination. The covariance is then maximized so that the correlation between υ_1 and $\nu_1 is$ maximized. This can be computed using the inner product of the score vectors $\hat{u}_1$ and $\hat{v}_1$: $\max\langle\hat{u}_1,\hat{v}_1\rangle=\rho_1^T \mathrm{AB}\gamma_1$ [13]. Applying the Lagrange multiplier method, the problem reduces to solving for the unit vectors ρ_1 and γ_1 so that γ_1 is maximized, and it is simply a matter of computing the eigenvalues of the matrix $M=A^T \mathrm{B}B^T \mathrm{A}$ with the eigenvectors. Model the regression of X on υ_1 and Y on ν_1, respectively [14]. Perform least squares estimation of the regression coefficients σ_1 and τ_1 by replacing A,B with the residual array andA_1 and B_1, and repeating the above steps until the absolute value of the elements in the residual array approximates to 0. Not performing this procedure once yields one σ_1 and τ_1. Repeating this procedure several times yields the partial least squares regression equation for Y $y_j=C_{j_1}X_1+C_{j_2}X_2+\cdots\ldots c_{j_m}X_m$ Multiple components are extracted for the above model, and the principal components are determined using a cross validity test to achieve higher accuracy.

3.1.1 Table of Factorial Variance Explained

Table 1. Table of Factorial Variance Explained

Latent Factors	X variance	Cumulative X variance	Y variance	Cumulative Y variance (R^2)	Adjusted R^2
1	0.428	0.428	0.883	0.883	0.878
2	0.34	0.768	0.032	0.915	0.907
3	0.13	0.898	0.009	0.924	0.912
4	0.086	0.984	0.035	0.959	0.95
5	0.016	1	0.002	0.961	0.949

The table above illustrates the cumulative explanatory power of potential factors, ranked by relevance. Here, the cumulative X variance signifies the information extraction from the independent variable, while the cumulative $Y^2(R^2)$ indicates information extraction from the dependent variable. These values serve as a foundation for establishing the maximum number of principal components for the parameter [11]. The factor variance explanation table results indicate that the top 3 potential factors collectively explain 80% of the independent variable's information, with the first potential factor explaining 80% of the information on its own.

3.1.2 Summary of the independent variable VIP (cumulative projected importance)

Table 2. Summary of the independent variable VIP

Variable	Factor1	Factor2	Factor3	Factor4	Factor5
Volume (shares)	0.574	0.688	0.685	0.694	0.694
P/E Ratio	1.225	1.204	1.2	1.194	1.194
SSE	1.321	1.307	1.303	1.285	1.284
Standard Deviation	0.525	0.524	0.526	0.6	0.605
Average	1.774	1.752	1.744	1.714	1.713
KDJ Indicator (K-value)	0.051	0.073	0.22	0.288	0.288
KDJ Indicator (D value)	0.001	0.124	0.163	0.239	0.247

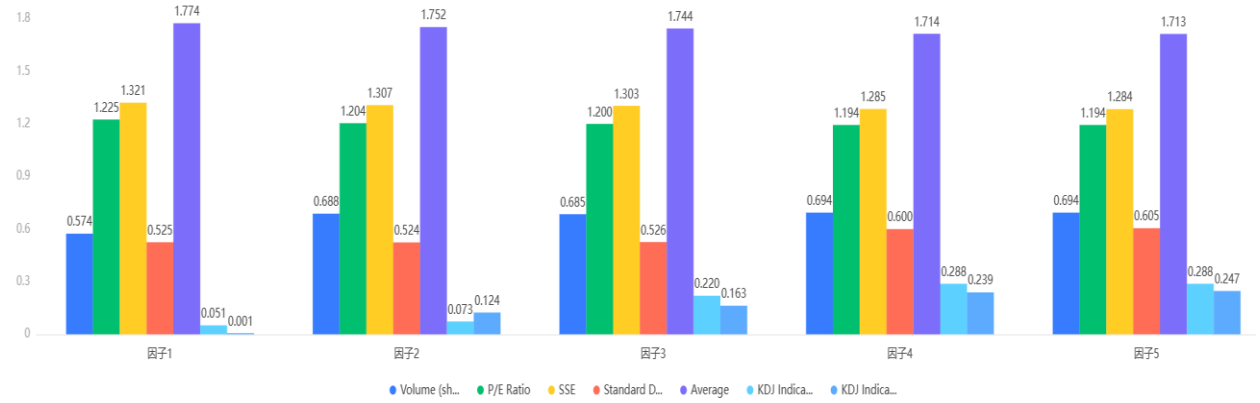

Figure 1. Cumulative Projected Importance

The table above displays VIP (Cumulative Projected Importance), which gauges the explanatory significance of X for Y under varying numbers of components. It can also guide us in determining the maximum number of principal components. The VIP of P/E Ratio, SSE, and Average is greater than 1, which means that it plays a greater role in explaining the underlying factors [12]. To provide a clearer sense of the explanatory power of each component, the VIP (Cumulative Projected Importance) is graphically presented in a bar chart.

3.1.3 Component matrix

Table 3. Component matrix

variant	factor1	factor2	factor3	factor4	factor5
Volume (shares)	0.217	0.772	-0.65	0.096	0.068
P/E Ratio	0.463	-0.151	0.29	0.2	-0.333
SSE	0.499	0.259	-0.452	0.564	0.006
Standard Deviation	-0.198	0.218	0.084	0.551	0.625
Average	0.67	-0.454	0.478	-0.487	0.392
KDJ Indicator (K-value)	0.019	0.106	0.722	-0.862	0.211
KDJ Indicator (D value)	0	0.251	0.231	0.195	-0.564
Closing price (yuan)	0.963	-0.294	0.114	-0.668	0.293

The table above provides the component matrix resulting from principal component analysis dimensionality reduction. This process condenses numerous indicators into a smaller number of principal components.

3.1.4 Table of factor loading coefficients

Table 4. Table of factor loading coefficients

variant	factor 1	factor 2	factor 3	factor 4	factor 5
Volume (shares)	0.312	0.729	-0.326	-0.332	-0.654
P/E Ratio	0.452	0.034	0.447	0.359	-1.285
SSE	0.537	0.127	-0.091	0.269	1.264
Standard Deviation	-0.175	0.367	0.649	0.686	0.448
Average	0.626	-0.265	0.041	-0.127	0.274
KDJ Indicator (K-value)	0.032	0.68	0.545	-0.363	0.546
KDJ Indicator (D value)	0.03	0.546	0.647	0.288	-0.19
Closing price (yuan)	0.618	-0.188	0.073	-0.428	0.188

The table above displays factor loading coefficients, which enable the assessment of the significance of hidden variables within each factor. Skillfully employing Principal Component Analysis (PCA), it's possible to convert high-dimensional data into a lower-dimensional form, thereby

improving the results of data analysis and machine learning tasks [13]. Concurrently, it can eliminate superfluous information, averting adverse effects on analysis and modeling.

3.1.5 Table of model coefficient results

Table 5. Table of model coefficient results

	Closing price (yuan)
variant	7.083
Volume (shares)	-0.026
P/E Ratio	0.056
SSE	-0.004
Standard Deviation	-0.082
Average	0.242
KDJ Indicator (K-value)	0.134
KDJ Indicator (D value)	-0.065

The table above displays the model coefficient results, which reveal the outcomes of the PLS model. These results primarily consist of model coefficients, utilized to assess the influence of independent variable X on dependent variable Y.

Thus, the standardized formula of the model is obtained as: y = 0.327 - 0.0*Volume (shares) + 0.233*P/E Ratio - 0.0*SSE + 0.86*Average + 0.009*KDJ Indicator (K-value) - 0.007*KDJ Indicator (D value) - 0.009*Standard Deviation.

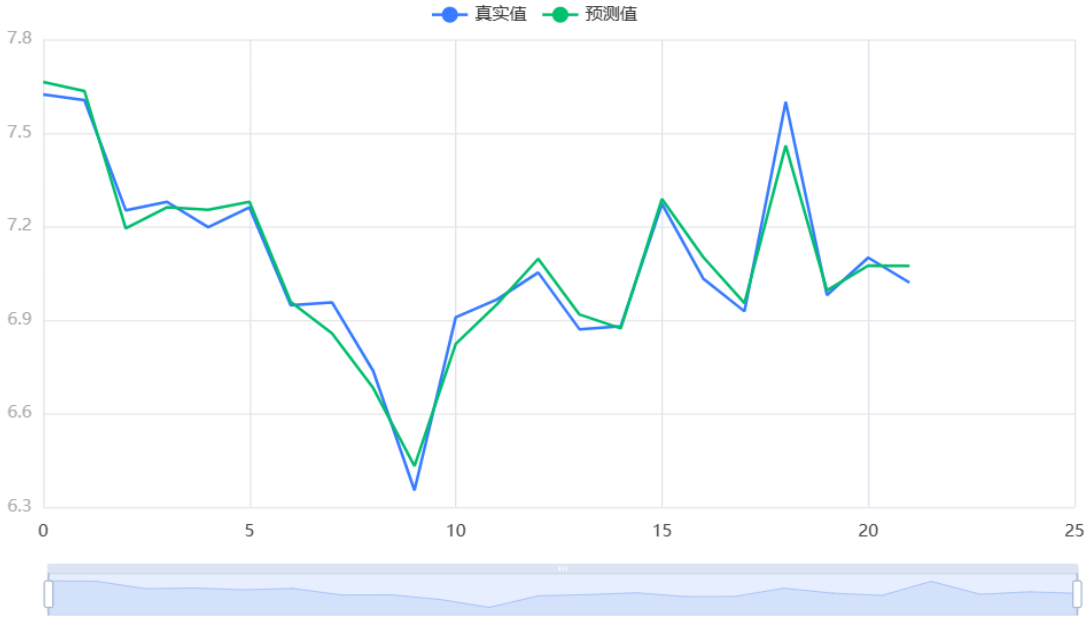

Figure 2. Plot of predicted versus real values

From the above figure, it can be seen that the predicted values of the partial least squares method are very close to the true values.

4. Limitations and Future Plans

PLS has significant potential applications in predicting stock market trends. PLS can be used in real-time market monitoring systems to detect risks and opportunities. Through integration with time series analysis, PLS aids investors in making faster decisions and adjusting their portfolios promptly. Additionally, PLS can integrate data from multiple sources, including market data, financial information, and sentiment analysis from social media. This integration offers a more comprehensive information foundation, facilitating a deeper comprehension of stock market intricacies [15]. Moreover, PLS can identify and quantify risk factors, assisting investors in more effective risk

management, a vital component for establishing robust risk management strategies. Nevertheless, it is crucial to recognize PLS's limitations. PLS performance is highly dependent on input data quality. Poor-quality, inaccurate, or incomplete data can result in imprecise model predictions, particularly in real-time market monitoring. Furthermore, PLS-generated models may become relatively complex, especially in high-dimensional datasets, potentially complicating model interpretation and maintenance. Additionally, PLS models are prone to overfitting, particularly when dealing with relatively small datasets. To address this concern, careful parameter tuning and cross-validation are essential. In future research, several exciting avenues warrant exploration. Firstly, the effective integration of deep learning methods, such as neural networks, with PLS can provide more robust models for predicting stock market trends. Deep learning excels in dealing with non-linear relationships and large datasets, making it a complementary tool to PLS. Furthermore, developing more robust real-time market monitoring tools by integrating PLS with other algorithms can improve the precision of market trends and risk predictions, benefiting both investors and market regulators. At a more profound level, integrating sentiment analysis with PLS can advance our comprehension of how market sentiment influences stock prices, ultimately enhancing predictive model accuracy.

5. Conclusion

When the observed data is insufficient compared to the number of predicted variables, and after selecting suitable indicators based on principal components relevant to the dependent variable [16], followed by their analysis and calculation using the correct formula, it is evident from the graph that the results obtained through the partial least squares method can effectively model the data for regression analysis. The results derived from these calculations exhibit a high level of precision. However, during the simulation test, the stock market was still in a recovery phase due to multiple factors influencing stock prices, including the recent impact of the epidemic. Unpredictable factors like epidemics were not considered, which may lead to disparities between experimental results and real-world outcomes. In the future, there is an expectation that the partial least squares algorithm can be enhanced further, reducing its constraints on variables to enhance prediction accuracy, especially in scenarios with limited data.

References

[1] Kao, L.J.; Chiu, C.C.; Lu, C.J.; Yang, J.L. Integration of nonlinear independent component analysis and support vector regression for stock price forecasting. Neurocomputing. 2013, 99, 534–542.

[2] Bai Lu. Simple prediction of stocks by partial least squares. Shenyang University of Technology, 2016.

[3] Kumbure, M.M.; Lohrmann, C.; Luukka, P.; Porras, J. Machine learning techniques and data for stock market forecasting: A literature review. Expert Syst. Appl. 2022, 197, 116659.

[4] Xiao Tian. China's Macroeconomic Data and Stock Returns- An Empirical Analysis Based on Partial Least Squares. Shanghai University, School of International Business and Management, 2006.

[5] Nianwu Deng, Hui Xu. Univariate partial least squares regression model and its application. Hubei: Wuhan University, 2001.

[6] Sun Chunyan, Chen Yaohui. Optimal stopping time analysis of American option pricing based on the least squares method'. Journal of Systems Engineering. 2003.11.

[7] Damien Cannavan Ronghong Huang a, Xiaowen Peng. Predicting stock returns with implied cost of capital: A partial least squares approach. Journal of Financial Markets. 2019.11

[8] Song, Y.; Lee, J.W.; Lee, J. A study on novel filtering and relationship between input features and target-vectors in a deep learning model for stock price prediction. Appl. Intell. 2019.

[9] Heng-Chang Zhang, Qing Wu, Fei-Yan Li, Hong Li. Multitask Learning Based on Least Squares Support Vector Regression for Stock Forecast. Axioms Journal.2022.11.

[10] Mohanty, D.K.; Parida, A.K.; Khuntia, S.S. Financial market prediction under deep learning framework using autoencoder and kernel extreme learning machine. Appl. Soft Comput. 2021.

[11] Kao, L.J.; Chiu, C.C.; Lu, C.J.; Yang, J.L. Integration of nonlinear independent component analysis and support vector regression for stock price forecasting. Neurocomputing. 2013.

[12] Cao, J.S.; Wang, J.H. Exploration of stock index change prediction model based on the combination of principal component analysis and artificial neural network. Soft Comput. 2020.

[13] Ceccato, C. D., & Lemgruber, E. F.. Valuation of American interest rate options by the Least-Squares Monte.Carlo method, Pesquisa Operacional, 2011.

[14] Abdel Sabour, S. A. and Poulin, R. Valuing real capital investments using the least squares Monte Carlo method.Engineering Economist, 2006. 51:141–160.

[15] Gourieroux, C. and Jasiak, J. Financial Econometrics: Problems, Models and Methods. Princeton Series in Finance.Princeton University Press, Princeton, NJ, first edition,2001.

[16] Cao, J.S.; Wang, J.H. Exploration of stock index change prediction model based on the combination of principal component analysis and artificial neural network. Soft Comput. 2020, 24, 7851–7860.