Stock Price Prediction Model Based on Generalized Gaussian Process Regression

Chenjia Ge[1, *, #], Shaohan Yu[2, #]

[1] School of Finance, Henan University of Finance and Economics and Law, Zhengzhou, China, 450000

[2]Business School of Hunan University, Changsha, China, 410082

*Corresponding author: 202234050415@stu.huel.edu.cn

#These authors contributed equally.

Abstract. Stock price prediction models hold significant research importance in the realm of quantitative trading. Given this, the present paper introduces a regression model based on the generalized Gaussian process. Specifically, two distinct models are proposed: one is a generalized Gaussian process regression model incorporating L1 regularization, and the other is a generalized Gaussian process regression model utilizing reproducing kernel representation. The former is capable of handling high-dimensional linear features and conducting crucial variable selection, while the latter can effectively model the nonlinear relationship between covariates and response variables. Moreover, both models are adept at capturing the temporal dependence inherent in stock price series. Under diverse simulation settings, the simulation results consistently demonstrate the strong applicability and competitiveness of the proposed methods. Ultimately, empirical analyses are carried out on stock price series data exhibiting varying volatility trends. The results substantiate the superior predictive performance and substantial application value of the proposed methods.

Keywords: Stock price prediction, Gaussian process, LASSO, regenerative kernel technique.

1. Introduction

In recent years, the stock market has garnered considerable attention due to its inherent volatility and complexity. Consequently, the quest for reasonable stock price prediction within a certain range has emerged as a popular research topic. Accurate stock price forecasting is of paramount importance as it not only empowers investors to effectively manage risks and avert losses, thereby enhancing their investment returns, but also furnishes enterprises with scientific financial decision-making support. Moreover, it aids governments and regulatory authorities in maintaining market stability and making timely adjustments to monetary and fiscal policies. Since the early 20th century, stock price forecasting methodologies have undergone significant evolution. They have transitioned from traditional technical and fundamental analysis to time-series regression models grounded in statistical modeling. Eventually, the advent of quantitative trading, which integrates computer technology and sophisticated mathematical models to facilitate trading decisions, has further advanced the field. Compared to traditional methods, quantitative trading has significantly enhanced market efficiency. This is attributed to its efficiency, discipline, and rapid responsiveness to market fluctuations, which substantially reduce the emotional interference of investors and lower the probability of human error. Additionally, quantitative trading offers valuable backtesting and optimization capabilities. Given these advantages, quantitative trading has proven to be particularly effective in navigating the complexities and volatility of modern financial markets. Building on the strengths of quantitative trading, this paper aims to refine the existing models to achieve more accurate stock price predictions.

Machine learning techniques have been extensively employed in the field of stock price forecasting. Early machine learning models were predominantly linear, positing that stock price changes could be elucidated through linear combinations of historical prices or other pertinent variables. Typical examples include autoregressive models (AR), moving average models (MA), and autoregressive moving average models (ARMA) [1]. These models capture the linear dependencies within stock price sequences but are constrained in their ability to utilize predictive information, making it

challenging to integrate additional predictors directly. In contrast, multifactor forecasting models exhibit greater versatility in practical applications. Ridge regression and LASSO regression [2-4] are frequently employed within this category to address high-dimensional linear predictor variables. However, linear models are notably inadequate when confronted with the nonlinear characteristics inherent in stock market data, as they are only capable of capturing linear relationships. With the ongoing evolution of machine learning, an increasing number of nonlinear regression models have been introduced to the realm of stock price prediction. For instance, random forest (RF) [5], support vector machines based on kernel functions (SVM) [6-7], k-nearest neighbors (KNN) [8], extreme gradient boosting (XGBoost) [9-10], and neural networks [11-13] have demonstrated proficiency in managing small samples and high-dimensional data while exhibiting robustness to outliers. Nevertheless, most nonlinear regression models face limitations in predictive performance when dealing with high-dimensional data, suffer from poor interpretability, and may experience slow convergence. Additionally, these methods often fail to adequately account for the autocorrelation effect present in stock price data, and their predictions typically yield a single value, devoid of any quantification of prediction uncertainty. In the domain of quantitative finance, the ability to provide such uncertainty quantification is of paramount importance.

This paper introduces a novel generalized Gaussian process regression model tailored to address the complexities of stock price prediction. The model encompasses two distinct formulations. The first is specifically designed for high-dimensional linear predictor variables and incorporates the L1 regularization technique [14-16]. This approach enables effective variable selection while efficiently handling high-dimensional problems. The second formulation leverages the regenerative kernel representation, which is adept at fitting complex structures with non-linear characteristics by transforming predictor variables in a non-linear manner. Both variants of the model assume that the relationship between predictor and response variables follows a Gaussian process a priori. Parameter estimation is achieved by minimizing a negative log-likelihood function criterion, augmented with penalties, to facilitate accurate stock price prediction. It is important to highlight that the prediction outcome of this model is not a singular value but rather a predictive distribution. This feature allows the model to provide robust support for the design of uncertainty quantification schemes in quantitative trading. Simulation experiments reveal that the generalized Gaussian process regression model based on L1 regularization not only delivers precise prediction results but also achieves efficient variable selection. Meanwhile, the model variant based on the regenerative kernel representation demonstrates exceptional prediction performance under non-linear function settings. Furthermore, the proposed method was rigorously tested using real-world data. The results confirm that the method maintains excellent prediction performance when applied to actual data. In summary, the generalized Gaussian process regression model presented in this paper exhibits strong applicability in the realm of stock price prediction. It effectively captures data characteristics and provides substantial support for uncertainty quantification in quantitative trading.

2. Theory and methodology

2.1. Generalized Gaussian Process Regression Model Based on L1 Regularization

Assume the relationship between the predictor variables and the response variable is given as follows:

$$Y = X^T\beta + g(s) + \varepsilon \tag{1}$$

where X represents the linear predictor variables; s denotes the predictor variables associated with random effects; β is the coefficient vector to be estimated; $g(s)$ represents an unknown random function; Y is a scalar response variable; and ε follows a standard normal distribution with mean 0 and variance 1. Assume that $g(s)$ follows a Gaussian process with a zero mean function and

covariance function defined by its kernel, i.e.,$g(s) \sim GP(0,k)$. Then, Y also follows a Gaussian process with mean function$X^T\beta$ and covariance function$k$, namely: $Y \sim GP(X^T\beta,k)$.

Given the training samples $\{x_i,s_i,y_i\}, i=1\ldots\ldots n$, the response variable data Y follows an n-dimensional normal distribution. The likelihood function is expressed as:

$$\frac{1}{\sqrt{(2\pi\sigma^2)^{\frac{n}{2}}|K|}}\exp[-\frac{(y_i-x_i^T\beta)K^{-1}(y_i-x_i^T\beta)}{2\sigma^2}] \tag{2}$$

where $K=[k(s_i,s_j),i,j=1\ldots\ldots n]$, and $|K|$ represents the determinant of K . Since the constructed marginal likelihood function is generally difficult to optimize directly, the negative log-likelihood is considered, resulting in the following loss function:

$$-\ln\{\frac{1}{\sqrt{(2\pi)^n}|K|}\exp[\frac{(y_i-x_i^T\beta)K^{-1}(y_i-x_i^T\beta)}{2\sigma^2}]\} \tag{3}$$

Considering the high dimensionality of the linear predictor variables, this study employs LASSO regression to apply L1 regularization on the coefficient β, leading to the following loss function:

$$-\ln\{\frac{1}{\sqrt{(2\pi)^n}|K|}\exp[\frac{(y_i-x_i^T\beta)K^{-1}(y_i-x_i^T\beta)}{2\sigma^2}]\}+\square\beta\square_1 \tag{4}$$

Based on the above, the coefficient β can be iteratively estimated using gradient optimization algorithms, with the estimated coefficient denoted as $\hat{\beta}$. For new test data x^*,s^*, the corresponding predictive distribution of y^* is given by:

$$y^*\square N[(x^*)^T\hat{\beta}+K^*K^{-1}(y_i-x_i^T\hat{\beta}),K^{**}-(K^*)^TK^{-1}(K^*)] \tag{5}$$

where $K^*=[k(s^*,s_i),i=1\ldots\ldots n]$, and$K^{**}=k(s^*,s^*)$. Finally, the mean of y^* is taken as the final predicted value. This method is referred to as the Generalized Gaussian Process Regression Model based on L1 Regularization (SGPR).

2.2. Generalized Gaussian Process Regression Model Based on L2 Regularization and Reproducing Kernel Representation

Assume the relationship between the predictor variables and the response variable is given as follows:

$$Y=f(x)+g(s)+\varepsilon \tag{6}$$

where $g(s)$ follows a Gaussian process, and since x is high-dimensional and its relationship with y is unknown, $f(x)$ can be expressed using the reproducing kernel expansion:

$$f(x)=\sum_{i=1}^{n}\alpha_i\,k(x_ix) \tag{7}$$

where $k(x_ix)$ is the reproducing kernel Hilbert space (RKHS), $i=1\ldots\ldots n$, and ε follows a standard normal distribution with mean 0 and variance 1. Given that $g(s)$ follows a Gaussian process with mean function zero and covariance function k_g, it satisfies the distribution:
$g(s)\sim GP(0,k_g)$. Thus, Y follows a generalized Gaussian process with mean function $\sum_{i=1}^{n}\alpha_i\,k(x_ix)$and covariance function k, namely:$Y\sim GP(\sum_{i=1}^{n}\alpha_i\,k(x_ix),k_g)$

Given the training samples $\{x_i,s_i,y_i\}, i=1\ldots\ldots n$, the likelihood function of the response data is expressed as:

$$\frac{1}{\sqrt{(2\pi\sigma^2)^{\frac{n}{2}}|K|}}\exp[-\frac{(y_i-\sum_{i=1}^{n}\alpha_i k(x_i x))K^{-1}(y_i-\sum_{i=1}^{n}\alpha_i k(x_i x))}{2\sigma^2}] \tag{8}$$

Taking the negative logarithm of the marginal likelihood function yields:

$$-\ln\{\frac{1}{\sqrt{(2\pi)^n|K|}}\exp[\frac{(y_i-\sum_{i=1}^{n}\alpha_i k(x_i x))K^{-1}(y_i-\sum_{i=1}^{n}\alpha_i k(x_i x))}{2\sigma^2}]\} \tag{9}$$

To account for the smoothness of the function, L2 regularization is applied to $f(x)$, leading to the following penalized negative log-likelihood function:

$$-\ln\{\frac{1}{\sqrt{(2\pi)^n|K|}}\exp[\frac{(y_i-\sum_{i=1}^{n}\alpha_i k(x_i x))K^{-1}(y_i-\sum_{i=1}^{n}\alpha_i k(x_i x))}{2\sigma^2}]\}+\square f \square^2 \tag{10}$$

By solving the above using gradient optimization algorithms, the kernel expansion coefficients can be obtained. For new test data x^*, s^*, the predicted distribution of y^* is given by:

$$y^* \square N[\sum_{i=1}^{n}\hat{\alpha}_i k(x_i x)+K^* K^{-1}(y_i-\sum_{i=1}^{n}\hat{\alpha}_i k(x_i x)), K^{**}-(K^*)^T K^{-1}(K^*)] \tag{11}$$

This method is referred to as the Generalized Gaussian Process Regression Model based on L2 Regularization and Reproducing Kernel Representation (KGPR).

3. Data analysis

3.1. Simulation experiment

To comprehensively evaluate the prediction performance of the proposed method, this paper conducts extensive simulation experiments by varying different model parameters. Specifically, the first set of experiments aims to verify the ability of the generalized Gaussian process regression model based on L1 regularization to handle dimensionality challenges and achieve effective feature selection. For this purpose, the connectivity between the predictor variables and the response variables is assumed as follows:

$$Y_i = X_i^T\beta + sin(S_i) + \xi_i, i = 1, 2, \cdots, n \tag{12}$$

We consider a scenario where the dimension of the linear predictor variable X is set to 5, and it follows a 5-dimensional standard normal distribution. Two cases are examined for the coefficient vector: one where all elements are zero, and another where the elements are non-zero, with non-zero values generated from a standard normal distribution. Additionally, the covariate S follows a standard normal distribution. The number of samples is set to n = 100, and the error term is assumed to follow a normal distribution. The dataset is randomly split into a training set and a test set in a 6:4 ratio. The training set is used to estimate the model parameters, while the test set is employed to evaluate the model's predictive performance. The results of variable selection are presented in Figures 1 and 2. Specifically, Figure 1 (left) corresponds to the case where the coefficient vector contains non-zero elements, while Figure 2 (left) pertains to the case where some coefficient vector elements are zero. Experimental results demonstrate that the proposed method effectively identifies important variables, thereby confirming its efficacy in feature selection. Moreover, Figures 1 (right) and 2 (right) illustrate the model's fitting performance on the test set under both scenarios. The prediction results indicate

that the proposed approach not only achieves high prediction accuracy but also exhibits robust performance in the presence of systematic errors. These findings suggest that the model is capable of performing both effective variable selection and accurate response prediction when handling complex and high-dimensional data. This validates the theoretical significance and practical applicability of the proposed method.

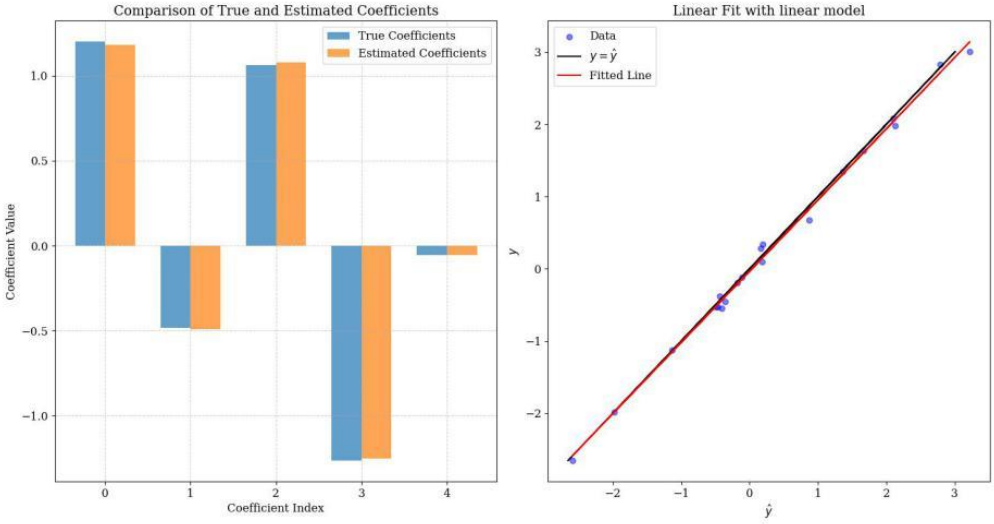


Figure 1 Feature selection and prediction results for the case where none of the coefficient vector elements is zero

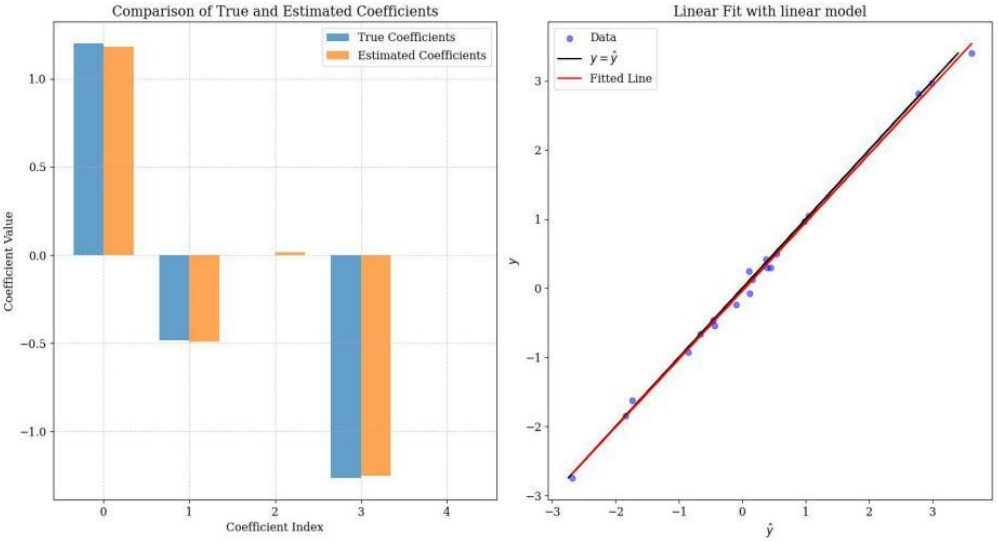


Figure 2 Feature selection and prediction results for the case where the coefficient vector element is partially 0

Then, in order to further verify the effectiveness of the proposed method in dealing with high-dimensional linear predictive variables, the prediction accuracies of the models in different dimensions are compared and analyzed in this paper. In order to objectively and comprehensively reflect the comprehensive performance of the proposed method, XGBoost, Random Forest (RF), Nonlinear Support Vector Machine (SVR), ElasticNet, and Gaussian Process Regression (GPR) are chosen as the comparison models. The Monte Carlo simulation method was used to calculate the evaluation indexes as follows: in the experiments, the training and test sets were randomly divided for a total of 10 times, and the mean square error (MSE) of the model on the test set was calculated under each division. Finally, the mean and standard deviation of the mean square error of the 10 experiments are used as the evaluation indexes. The smaller the mean value of the mean square error, the higher the prediction accuracy of the model; the smaller the standard deviation, the more robust the model is. The experimental results are shown in Table 1. From the table, it can be clearly seen

that the proposed method exhibits the highest prediction accuracy and the strongest robustness in different dimensions. This indicates that the proposed method effectively handles the problem of nonlinear relationship modelling and dimensionality catastrophe by introducing the L1 regularization technique and assuming that the nonlinear relationship obeys the Gaussian process a priori. The experimental results fully demonstrate the significant advantages and reliability of the proposed method in high-dimensional linear predictor variable scenarios.

Table.1. Mean and variance of mean square error for each model at different feature dimensions

Methods	D=5	D=10	D=15	D=20
SGPR	0.062(0.059)	0.013(0.002)	0.020(0.011)	0.020(0.004)
XGBoost	0.071(0.028)	2.001(0.619)	4.801(0.431)	13.529(3.971)
RF	0.062(0.026)	2.014(0.737)	4.156(1.308)	10.041(2.847)
SVR	0.387(0.062)	2.431(0.644)	4.823(1.484)	13.723(6.056)
ElasticNet	0.507(0.085)	3.169(0.947)	4.092(1.442)	7.212(3.351)
GPR	0.491(0.078)	1.658(0.491)	2.459(0.913)	6.489(3.206)

Second, this paper presents a convergence analysis of the generalized Gaussian process regression model based on the L1 penalty. The specific settings are such that the sample sizes n are 50, 100, 200 and 500, and the rest of the parameters are kept consistent with the initial experimental settings. The prediction performance of the model is evaluated by Monte Carlo simulation method for different sample sizes. The experimental results are shown in Table 2. It can be clearly seen from the table that with the increase of sample size, the prediction accuracy of the proposed method on the test set gradually improves, and the mean of mean square error (MSE) decreases significantly, while the standard deviation also decreases gradually, which indicates that the robustness of the model is enhanced. From the experimental results, it can be concluded that the proposed method is able to show higher prediction accuracy and stronger robustness when the sample size increases. This verifies the convergence of the proposed method from the experimental point of view, i.e., the increase of sample size can effectively improve the performance of the model, which further proves the reliability and applicability of the method in practical applications.

Table.2. Mean and variance of mean square error for each model with different number of samples

Methods	n=50	n=100	n=200	n=500
SGPR	0.065(0.095)	0.013(0.002)	0.010(0.0013)	0.00997(0.001)
XGBoost	1.44(0.499)	2.001(0.619)	1.677(0.348)	1.471(0.317)
RF	0.999(0.323)	2.014(0.737)	1.581(0.321)	1.758(0.216)
SVR	1.007(0.312)	2.431(0.644)	1.441(0.440)	1.441(0.440)
ElasticNet	1.121(0.323)	3.169(0.947)	2.409(0.478)	3.357(0.654)
GPR	0.832(0.256)	1.658(0.491)	0.803(0.140)	0.712(0.159)

Then, the paper further explores the effect of different dimensions of the covariate S and different nonlinear functional forms on the proposed method. The rest of the parameter settings in the experiment are kept consistent with the initial experiment. The experimental results are shown in Tables 3 and 4. From the tables, it can be seen that the proposed method performs well in dealing with the modelling of random effects between the covariates S and the response variable. When the dimensionality of the covariates increases, although the prediction accuracy slightly decreases, the overall performance remains stable, indicating that the proposed method is robust to high-dimensional covariates. In addition, the proposed method is still able to accurately capture the nonlinear relationship between the variables under different nonlinear functions, and the prediction performance maintains a high level. In summary, the experimental results show that the proposed method can not only effectively handle the random effects between the covariates and the response variables, but also maintain a good prediction performance when the dimensionality of the covariates

increases. This property proves the flexibility and applicability of the proposed method in dealing with complex data relationships.

Table.3. Mean and variance of mean square error for each model in covariate dimensions

Methods	S=50	S=100	S=200	S=300
SGPR	0.065(0.095)	0.013(0.002)	0.010(0.0014)	0.011(0.0015)
XGBoost	1.437(0.499)	2.001(0.619)	1.677(0.348)	0.787(0.0935)
RF	0.999(0.323)	2.014(0.737)	1.581(0.322)	0.873(0.134)
SVR	1.007(0.312)	2.431(0.644)	1.304(0.260)	1.254(0.353)
ElasticNet	1.121(0.323)	3.169(0.947)	2.409(0.478)	2.053(0.494)
GPR	0.832(0.256)	1.658(0.491)	0.803(0.140)	0.634(0.146)

Table.4. Mean and variance of mean square error for each model with different function structures

Methods	$g(S)=\sin(\frac{2}{3}\pi S)$	$g(S)=\cos\frac{2}{3}\pi S$	$g(S)=e^{S}$	$g(S)=\sin S+S^{2}$
SGPR	0.0126(0.0021)	0.0126(0.0022)	0.012(0.00196)	0.013(0.0045)
XGBoost	2.001(0.619)	2.432(0.502)	2.329(0.738)	3.215(2.077)
RF	2.014(0.737)	2.202(0.557)	2.118(0.796)	2.999(1.860)
SVR	2.431(0.644)	2.313(0.623)	2.379(0.823)	3.353(1.715)
ElasticNet	3.169(0.947)	3.261(0.767)	2.956(1.226)	4.040(1.901)
GPR	1.658(0.491)	1.590(0.421)	1.421(0.485)	2.782(1.718)

Finally, we examine the performance of the generalized Gaussian process regression model based on L2 regularization and regenerative kernel representation. The hypothesized predictor variables are connected to the response variables as shown below:

$$Y_i = 0.1 * (X_{1i} + X_{2i})^3 + sin(\frac{3}{5}\pi S_i) + \xi_i \tag{13}$$

where both X1i and X2i obey the standard normal distribution and Si obeys the standard normal distribution. In this experiment, the error terms obey the standard normal distribution and are independent of the predictor variables. The number of samples is set to 100, and the dataset is divided into a training set and a test set in the ratio of 6:4, where the training set is used to estimate the model parameters and the test set is used to test the predictive performance of the model. In order to ensure the optimal performance of the model, the hyperparameters of the proposed method are determined by a cross-validation method. The prediction results on the test set are shown in Fig. 3. From the results, it can be seen that when there are nonlinear relationships between the predictor variables and the response variables, the proposed method is able to effectively capture and fit these nonlinear relationships by selecting appropriate kernel functions. In addition, the fitted curves are highly consistent with the true values, indicating that the proposed method has high prediction accuracy and strong nonlinear modelling capability. This indicates that the proposed method can still show excellent adaptability and robustness in complex nonlinear situations.

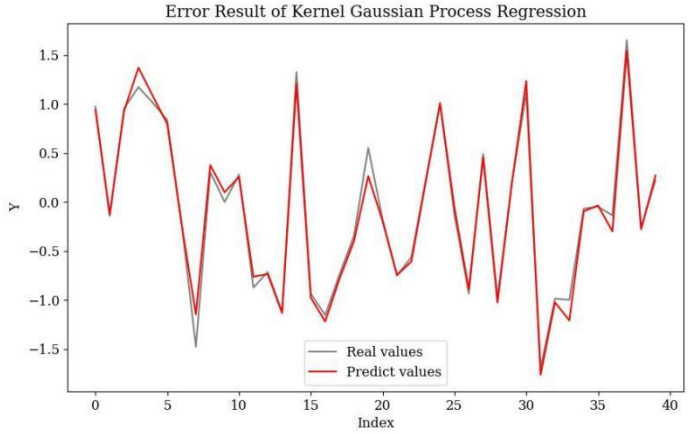


Figure 3 Fitting of the nonlinear function on the test set

In order to further validate the predictive performance and applicability of the generalized Gaussian process regression model based on L2 regularization and regenerative kernel representation, a series of experiments are designed in this paper to investigate the performance of the method under different sample numbers and different connection functions. In the experiments, the other parameter settings are consistent with the initial experiments, and the specific experimental results are shown in Tables 5 and 6. From the experimental results, it can be seen that the proposed method can accurately fit nonlinear connection functions of different complexity while dealing with the random effects, and shows strong prediction performance. In addition, with the increase of the number of samples, the prediction accuracy of the method is further improved and the robustness is significantly enhanced. This indicates that the proposed method is not only able to effectively capture the nonlinear relationship, but also exhibits good adaptability and convergence under different sample sizes.

Table.5. Mean and variance of mean square error for each model with different number of samples

Methods	n=50	n=100	n=200	n=500
KGPR	0.273(0.546)	0.237(0.484)	0.155(0.253)	0.128(0.206)
XGBoost	0.364(0.298)	0.745(0.684)	0.479(0.344)	0.224(0.126)
RF	0.343(0.330)	0.965(0.855)	0.626(0.469)	0.375(0.212)
SVR	0.257(0.301)	1.224(1.203)	0.527(0.497)	0.268(0.168)
ElasticNet	0.643(0.354)	2.297(1.459)	1.817(1.129)	1.456(0.346)
GPR	0.515(0.290)	1.171(1.160)	0.507(0.469)	0.252(0.140)

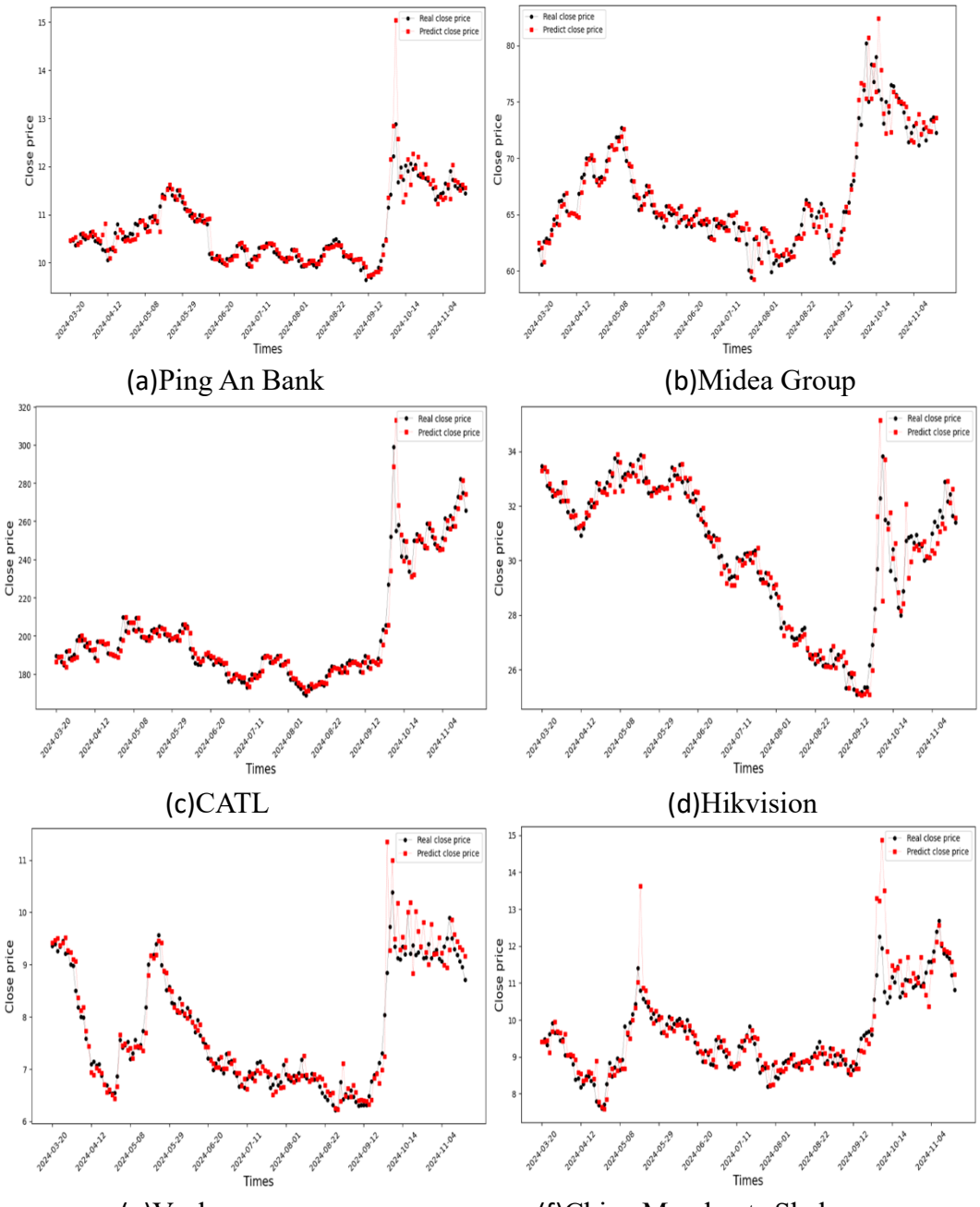
Table.6. Mean and variance of mean square error for each model with nonlinear function structures

Methods	$g(X_1, X_2) = {X_1}^2 + {X_2}^2$	$g(X_1, X_2) = sin X_1 \, cos X_2$	$g(X_1, X_2) = e^{X_1 + X_2}$	$g(X_1, X_2) = {X_1}^3 + {X_2}^3$
KGPR	0.096(0.676)	0.015(0.011)	0.054(0.256)	0.035(0.132)
XGBoost	1.314(0.754)	0.156(0.041)	1.567(1.187)	6.537(5.156)
RF	1.418(1.379)	0.158(0.042)	1.134(0.731)	5.546(4.004)
SVR	1.191(1.987)	0.095(0.047)	1.138(0.832)	18.53(10.447)
ElasticNet	4.343(2.947)	0.585(0.201)	2.833(1.450)	13.663(8.207)
GPR	1.807(2.527)	0.193(0.070)	1.227(0.727)	10.697(7.437)

3.2. Analysis of actual data

Equity forecasting is crucial in quantitative investing, helping investors to optimize their portfolios, reduce risk and increase returns through effective modelling and data analysis. At the same time,

forecasting models can dynamically adapt to market changes, thus enabling investors to maintain a competitive advantage in a complex and volatile market environment. In order to verify whether the model proposed in this paper can have the same accuracy and robustness as simulation experiments when dealing with real data, and whether it can be applied to the analysis of real data, the experiments of real data analysis are conducted here. The experimental data come from the Tushare website, and six listed companies, namely Ping An Bank, Midea Group, Ningde Times, Hikvision, Vanke and China Merchants Shekou, are selected, and the share price data of these six companies from 2 January 2024 to 29 November 2024 are extracted for analysis. These six companies come from different industries and are very representative of their respective industries, so the selection of these six companies for analysis can illustrate the broad applicability and effectiveness of the model in this paper. This part of the experimental data analysis adopts the rolling sliding method of prediction principle, by moving a fixed-size window on the time series data, gradually updating the training data and retraining the model, so as to dynamically adapt to the data changes and improve the accuracy of prediction. The hyperparameters for prediction are determined by cross-validation.



(a)Ping An Bank (b)Midea Group

(c)CATL (d)Hikvision

(e)Vanke (f)China Merchants Shekou

Figure 4 The result of actual data analysis for six listed companies

In the figure, the black line graph is plotted based on the real closing price, and the red line graph is plotted based on the predicted closing price of this model. As shown in the figure, the error between the predicted closing price and the real closing price in the real data experiments of the above six enterprises is small, and the similarity between the fitted closing price curve and the real closing price curve is extremely high, so this model has a high prediction accuracy. At the same time, based on the analysis of the real data of six enterprises, the model shows excellent prediction ability for all six enterprises, which proves that the model has wide applicability in the real application.

4. Conclusion

This study proposes a model based on generalized Gaussian process regression, which is capable of performing variable selection, addressing the random effects of the response variable, and capturing the nonlinear relationships between covariates and the response variable. The model effectively resolves challenges such as selecting influencing factors in stock price forecasting, processing complex stock market data, and addressing autocorrelation within stock price series. Compared with other models, the proposed approach not only provides point forecasts of stock prices but also offers uncertainty intervals for the forecasts. Simulation experiments demonstrate that the proposed model outperforms classical models such as Random Forest, Nonlinear Support Vector Machines, Gaussian Process Regression, and XGBoost, highlighting its capability for efficient variable selection. Furthermore, the favorable performance observed in real-world data experiments substantiates the model's high applicability and predictive performance.

The application of this model can be extended to stock price volatility prediction, risk management, and trading signal generation. For instance, by leveraging the model to predict future stock price movements and formulating corresponding trading strategies based on prediction intervals, investors can optimize returns by adopting short positions when a price decline is anticipated and long positions when prices are expected to rise, thereby achieving risk-controlled investment decisions. Additionally, the model's applicability extends beyond financial markets to various fields such as weather forecasting. In this context, the model can capture dynamic characteristics of temperature, weather, and climate data, filter relevant influencing factors, and generate reliable forecasts. To further enhance the model's generalizability and predictive capabilities, the selection of kernel functions and hyperparameter optimization are crucial. These factors significantly impact the model's fitting and generalization performance. The choice of kernel function should be tailored to the specific characteristics of the data, while hyperparameter optimization can be conducted using techniques such as grid search, random search, and Bayesian optimization.

References

[1] Lai Huihui. Forecast of output VAT based on ARMA model under the background of big data[J].Tax Research,2019,(02):41-46.

[2] Li Bin, Shao Xinyue, Li Yueyang. China Industrial Economics,2019,(08):61-79.

[3] Mao Jie, Chen Mizhou, Xu Lei, et al. Research on the marginal effectiveness of pricing factors in China's stock market based on double choice LASSO model[J].Systems Engineering - Theory & Practice,2024,44(09):2993-3008.

[4] Feng G, Giglio S, Xiu D. Taming the factor zoo: A test of new factors[J]. The Journal of Finance, 2020, 75(3):1327–1370.

[5] Fang Yi, Chen Yuzhi, Wei Jian. Artificial intelligence and China's stock market: A quantitative research on portfolio based on machine learning prediction[J].Industrial Technology Economics,2022,41(08):83-91.

[6] Lin Mingsong, Yang Xiaomei, Yang Zhixia. Application of structured maximally spaced dual support vector machine in stock forecasting[J].Computer Engineering and Applications,2024,60(11):346-355.

[7] Ma Tingting, Yang Zhixia, Ye Junyou. Robust biparametric interval support vector machine[J].Computer Engineering and Applications,2022,58(09):74-82.

[8] Wang Bingyu, Liu Yongjun. Research on stock price prediction based on signal-to-noise ratio based on KPCA-SVM-KNN algorithm[J].Computer and Digital Engineering,2022,50(04):685-690.

[9] Gou Xiaoju, Wang Qian. Research on stock return direction based on data mining technology[J].Operations Research and Management,2021,30(01):163-169.

[10] Wang Yan, Guo Yuankai. Application of improved XGBoost model in stock forecasting[J].Computer Engineering and Applications,2019,55(20):202-207.

[11] Yu Zhuoxi, Qin Lu, Zhao Zhiwen, et al. Stock price prediction based on principal component analysis and generalized regression neural network[J].Statistics and Decision,2018,34(18):168-171.

[12] Hu Qingyu, Chen Qi'an. Stock Market Return Prediction Based on Prior Feedforward Neural Network[J/OL].Systems Engineering Theory & Practice,1-23[2025-01-24].

[13] Kong Yinying, Ke Ruikai, Hu Yamei, et al. Stock price prediction model based on one-way adaptive graph neural network in each sector of stocks[J].Journal of South China Normal University(Natural Science Edition),2023,55(04):100-107.

[14] Wang Man, Zhang Miaomiao. Volatility and stock price prediction considering high-dimensional macro information[J].Statistics and Decision,2022,38(20):138-143.

[15] Lu Yuhong, Song Jiali, Wang Meng, et al. Research on prediction model based on deep neural network fusion sparse grouping lasso[J].China Health Statistics,2021,38(06):821-827.

[16] Hu Yanmei, Yang Bo, Duobin. Regularized logistic regression based on network structure[J].Computer Science,2021,48(07):281-291.