Stock Price Forecasting Model Research Based on Multi-Kernel Gaussian Process Regression and Stacking Algorithm

Ruiling Liu *, Xuyang Sun

School of Finance, Southwest University of Finance and Economics, Chengdu, China, 611130

* Corresponding Author Email: xxxkwee@163.com

Abstract. With the rapid development of data collection and storage technologies, the variety of non-Euclidean predictors in the field of quantitative finance is increasing, and how to integrate them with the information of the traditionally used Euclidean factors to improve the performance of predictive models is an important research issue. Based on this, this paper proposes an integrated learning model based on multi kernel learning and Gaussian process regression. Specifically, on the one hand, the proposed method integrates the prediction information from different sources by using different kernel functions as a measure of the difference between elements in the Euclidean and non-Euclidean spaces and based on the Gaussian process regression model. On the other hand, the method improves the prediction accuracy and overall robustness of the prediction model by using the stacking algorithm to quadratically fit the output of the Gaussian process regression model as a new input variable. In addition, the learner used in the secondary fitting is model-free. Simulation analyses illustrate the effectiveness of the proposed method under different experimental settings. Finally, by predicting real stock price data, the results show that the proposed method has a smaller prediction error compared to some classical prediction models.

Keywords: Stock price forecasting, Gaussian process regression, Ensemble learning, Uncertainty estimation.

1. Introduction

Effective stock price forecasting provides investors with a window into market movements and helps identify potential investment opportunities and risks. Currently, stock price forecasting methods are divided into two main categories: forecasting models based on time-series information and forecasting models based on factor information. Models based on time-series information: this type of model usually relies on historical stock price data for forecasting, such as ARIMA model [1] and LSTM neural network model [2]. However, relying only on time-series information may not adequately capture future stock price volatility trends, resulting in poor predictions.

Models based on factor information: multifactor prediction models make predictions by considering more factors that affect the stock price. Linear regression models assume a linear relationship between the stock price and a set of influencing factors and predict the stock price by finding the best linear combination of these factors [3]. But it is not flexible enough to face the non-linear characteristics of financial markets. The volatility of stock prices is often affected by multiple factors, which have complex interactions and nonlinear relationships that are beyond the handling capacity of traditional linear regression models. In addition, with the increase in the number of factors, linear regression models are prone to fall into the dimensionality catastrophe problem [4]. To solve these problems, regularisation techniques provide effective solutions, especially LASSO regression [5] and ridge regression [6], which improve the stability and predictive ability of the model under high-dimensional data by introducing penalty terms. In response to the possible non-linear relationship between stock prices and factors, non-parametric models deal with complex relationship structures through more flexible designs. For example, kernel regression models use kernel methods to map the data into a high-dimensional space, thus capturing complex nonlinear patterns. Support Vector Machines (SVMs) based on kernel functions excel in classification and regression problems when dealing with small samples and high-dimensional data [7]. Decision trees construct models by recursively partitioning the dataset, and their ease of understanding and implementation have made them popular in practical applications [8]. In addition, integrated learning models such as Random

Forest [9] and XGBoost [10] improve prediction accuracy and stability by integrating multiple decision trees, and these models show significant advantages in dealing with nonlinear relationships. However, non-parametric models converge slowly when dealing with high-dimensional problems, which affects prediction accuracy.

In the field of financial quantification, the accuracy of prediction models is an important research goal, and the quantitative estimation of uncertainty in prediction results is also a key direction. Through uncertainty quantification, the prediction results become more reliable and provide more detailed technical support for decision makers. With the advancement of data collection and storage technology, the introduction of non-Euclidean geometric factors provides a new research perspective for stock price prediction models. Traditional models are usually based on Euclidean spatial assumptions, especially when dealing with sentiment factors, and tend to ignore the intrinsic geometric structure of the data, resulting in compromised prediction accuracy. Unstructured data (e.g., text data and social media sentiment) contain rich market information that directly or indirectly affects stock prices and reveals subtle changes in market sentiment and investor behaviour, which are difficult to capture by traditional models [11-13].

To address the limitations of existing models and leverage multiple methods, this paper introduces an innovative integrated learning model. Combining multi-kernel learning with Gaussian process regression (GPR), it integrates information from both Euclidean and non-Euclidean factors. Through refined kernel functions, the model captures and measures differences in these spaces, using GPR to integrate the information. To further improve accuracy and robustness, a stacking algorithm is applied, learning GPR outputs as new features in two stages. This flexible approach, not limited by specific learners, is validated through simulation and real data analysis across various conditions.

2. Theory and Methods

2.1. Gaussian Process Regression

Gaussian Process Regression (GPR) is a non-parametric Bayesian regression method for solving regression problems. The core idea of GPR is to model the similarity of the input data through the covariance function (i.e., the kernel function) and infer the output of the unknown data based on the multivariate Gaussian distribution. Compared to traditional regression methods, GPR not only provides a predicted mean for the unknown data but is also able to provide an estimate of the uncertainty through the covariance matrix. For a given training dataset and corresponding target output, GPR assumes that the outputs of these data points obey a joint Gaussian distribution, which is denoted as:

where is the covariance matrix, computed from the kernel function, which is used to measure the similarity between data points. is the variance of the observation noise and I is the unit matrix. When a new input point is available, the goal of GPR is to infer the conditional distribution of the outputfor that point. The conditional distribution of the new target value still obeys a Gaussian distribution according to Bayes' formula:

where the predicted mean and the predicted covariance are given by the following equations, respective

In this model, the choice of kernel function has a crucial impact on the performance of GPR. The kernel function not only determines the way of similarity measure between data points but also affects the smoothness and prediction accuracy of the model. In this study, in order to adequately handle different types of data, we chose two different kernel functions to deal with Euclidean and non-Euclidean data respectively: the RBF kernel function (Radial Basis Function) is used for Euclidean data. This kernel function is commonly used in traditional regression problems and has strong smoothing and global properties. The cosine similarity kernel function will be used to process the non-Euclidean data. This kernel function measures the angular relationship between data points by cosine similarity and is more suitable for processing vectorised non-Euclidean data. Its formula is:

where denotes the cosine similarity between the two vectors and is a parameter used to adjust for the effect of similarity. This dual kernel function strategy allows us to exploit both the local and global structure of the data when dealing with complex data, thus improving the generalisation ability and prediction performance of the GPR model.

2.2. Stacking algorithm based on multicore learning and Gaussian process regression model

In order to further improve the prediction performance of the regression task and reduce the uncertainty caused by a single model, we propose an integrated model based on multicore learning and Gaussian process regression, IGPR (Integrated Gaussian Process Regression). IGPR combines the flexibility of multikernel learning and the advantages of the Stacking integration strategy and performs well in dealing with complex heterogeneous data. Suppose the input dataset X has dimensional features and contains samples. In order to capture the properties of different feature types, we classify the input features into Euclidean feature and non-Euclidean feature with dimensions, respectively. denoted as For the Euclidean feature, we use the RBF kernel function to model the similarity between the data:

For the non-Euclidean feature we instead use the cosine similarity kernel function to capture the nonlinear structure of the data

In the IGPR model, we train a separate Gaussian process regression model for each subset. These two models are constructed based on the RBF kernel function and the cosine similarity kernel function, respectively, and the corresponding covariance matrices are generated:

IGPR uses the Stacking integration strategy to combine the outputs of multiple Gaussian process regression models based on different kernel functions. Assuming that we havebase learners, each of which predicts, we stack these into a new feature matrix as follows:

and passed as input to the second layer (ridge regression model) for final fitting and prediction. The final prediction of Stacking integration is calculated by ridge regression.

3. Data Analysis

3.1. Data Generation and Experimental Setup

In this experiment, a synthetic dataset is designed and generated in order to verify the validity of the proposed IGPR model. The feature matrix X contains two parts: the Euclidean feature and the non-Euclidean feature Among them, the Euclidean feature: randomly generated from the uniform distribution, represents the traditional numerical data in the financial market. Specifically, each column of data in the Euclidean Feature Matrix is randomly generated from an independent uniform distribution with the range set to to ensure that the data is uniformly distributed within a certain range. For the non-Euclidean feature it is generated from the normal distribution which is subsequently mapped to the non-Euclidean geometric space by a non-linear transformation. In order to model more complex financial data relationships, each column of data for the non-Euclidean feature is first sampled from the normal distribution and then a cosine similarity transformation is applied to capture the non-linear structure of the data. The target variable Y is generated by the following link function:

where:and γ are coefficients adjusting the contribution of Euclidean and non-Euclidean features to the target variable; β and δ are weight vectors multiplied by the Euclidean feature and the non-Euclidean feature respectively; and is a noise term sampled from the normal distribution to model the random errors in the data. With the above connection function, the target variable Y combines the effects of both Euclidean and non-Euclidean characteristics, reflecting the complex linear and non-linear relationships that may exist in financial markets. In order to test the performance of the IGPR model under different data complexity, we designed the following experimental setups: firstly, **the number of features** setup: the first case is a 3-dimensional feature: and are 3-dimensional data, respectively; this setup simulates a simpler regression problem to test the model's performance on low-dimensional data. The second case is 6-dimensional features: and are 6-dimensional data, respectively. This setting increases the complexity of the data and simulates a financial data prediction

scenario with high-dimensional features. The second is **the number of simulations** setting: the first case is 10 simulations: 10 independent simulations are performed for each feature number setting to assess the average performance and stability of the model. The second case is 20 simulations: 20 independent simulations for each feature number setting to further validate the robustness of the model and to observe the model's performance with larger amounts of data. In each experiment, the generated feature matrix X and the target variable Y are used as the training and testing datasets for the model. We use the mean square error (MSE) as the main evaluation metric to compare the performance of the IGPR model with other kernel function Gaussian process regression models under different experimental conditions.

In order to analyse the empirical data, the study selects three representative stocks (codes 0000019, 0000521 and 0000728, respectively) for the purpose of conducting experiments to test the predictive ability of the proposed model under different experimental conditions. The data are obtained from the TUShare stock price data website. In particular, the three selected stocks are drawn from disparate industries, thereby ensuring a more heterogeneous data set and a superior ability to test the adaptability of the model in diverse market environments. The data set for each stock comprises multiple trading days and includes a range of multi-dimensional characteristics, such as opening price, closing price, high price, low price and volume. In order to streamline the analysis and concentrate on a comparison of the model's performance, the primary prediction target is selected to be the stock's daily closing price gain. In order to select non-European characteristics, the Market Transaction Rate (TURN) and the Consumer Confidence Index (CCI) are employed. The Market Transaction Rate reflects the trading activity of investors over a given period and may therefore be considered an indicator of the willingness of investors to trade in the market, as well as fluctuations in their emotions. These, in turn, affect short-term fluctuations in stock prices. The Consumer Confidence Index (CCI) is employed to evaluate consumers' perceptions of the prevailing economic circumstances and their anticipations for the prospective economic outlook. This index offers a comprehensive representation of consumers' subjective sentiments pertaining to the economic situation, income expectations, and psychological state of consumption.

In order to fully evaluate the performance of the model, firstly, we adopt the common time series segmentation method in the division of the training and test sets. Specifically, for each stock's data, we use the first 80% of the data in chronological order as the training set and the last 20% as the test set. The purpose of this division is to simulate the prediction task in real application scenarios, i.e., the model is trained on known historical data and predicted on future data. The method is able to test the generalisation ability of the model while ensuring that the test set data does not leak into the training process. We conducted 10, 20 and 40 experiments on each stock. The number of experiments was chosen to explore the differences in model performance at different sample sizes. For each round of experiments, we repeat them several times to ensure the reliability and stability of the results. Specifically, 10 experiments correspond to scenarios with small sample sizes, which may occur in real applications with scarce data or real-time prediction; 20 and 40 experiments correspond to scenarios with medium and large sample sizes, respectively, which simulate scenarios with long-term data accumulation or large-scale data applications. The evaluation metric used in the experiments is the mean square error (RMSE). RMSE is a standard indicator of how much the predicted value deviates from the true value, and the smaller the value, the higher the predictive accuracy of the model.

3.2. Comparison of Experimental Results

Firstly, Figure 1 summarises the average performance of all models under different number of features and number of simulations. Each bar represents the mean value of the model under the corresponding condition, revealing the relative performance of each model under different experimental environments. The IGPR model maintains a low mean square error under almost all experimental conditions, showing superior prediction performance.

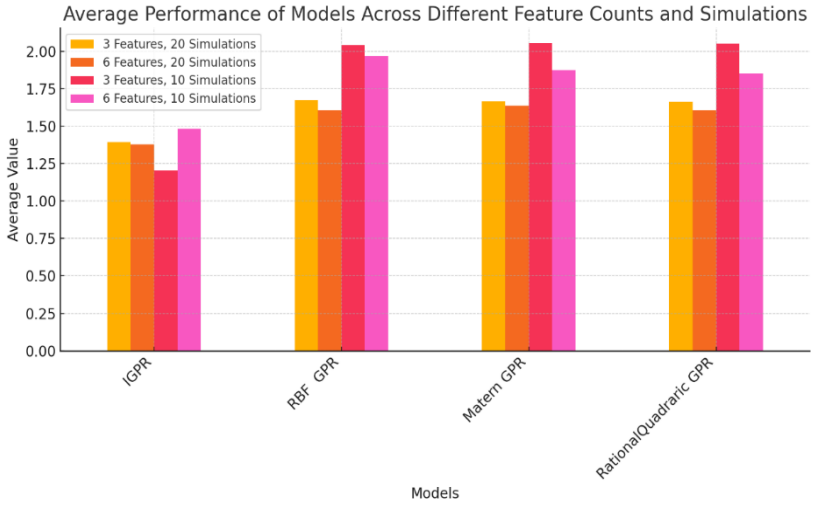

Figure 1: Average performance of each model in different experimental settings

As illustrated in Figure 1, the IGPR model demonstrates the lowest mean square error (MSE) at feature numbers 3 and 6 with a simulation count of 20. This suggests that the model is effective in addressing complex feature spaces and large data sets. In contrast, the RBF, Matern and RationalQuadratic GPR models demonstrated a reduction in prediction accuracy when confronted with an increased number of features and a more intricate data set. Additionally, these models exhibited a notable increase in error variability, particularly when the number of features was augmented. To gain further insight into the prediction error distributions of the different models, Figure. 2 depicts the error distributions of each model under the four experimental settings via box-and-line plots. The box-and-line plots provide comprehensive information about the median error, quartile differences, and outliers.

Furthermore, an additional verification of the superiority of the IGPR model can be obtained by analysing the summary statistics presented in Table 1. The mean, standard deviation, minimum and maximum values of each model in the table provide further evidence that the IGPR model consistently outperforms the other models in terms of prediction performance under each experimental condition. Both from the perspective of the mean and the standard deviation, the IGPR model maintains relatively small error fluctuations, which provides strong evidence of its stability and prediction accuracy.

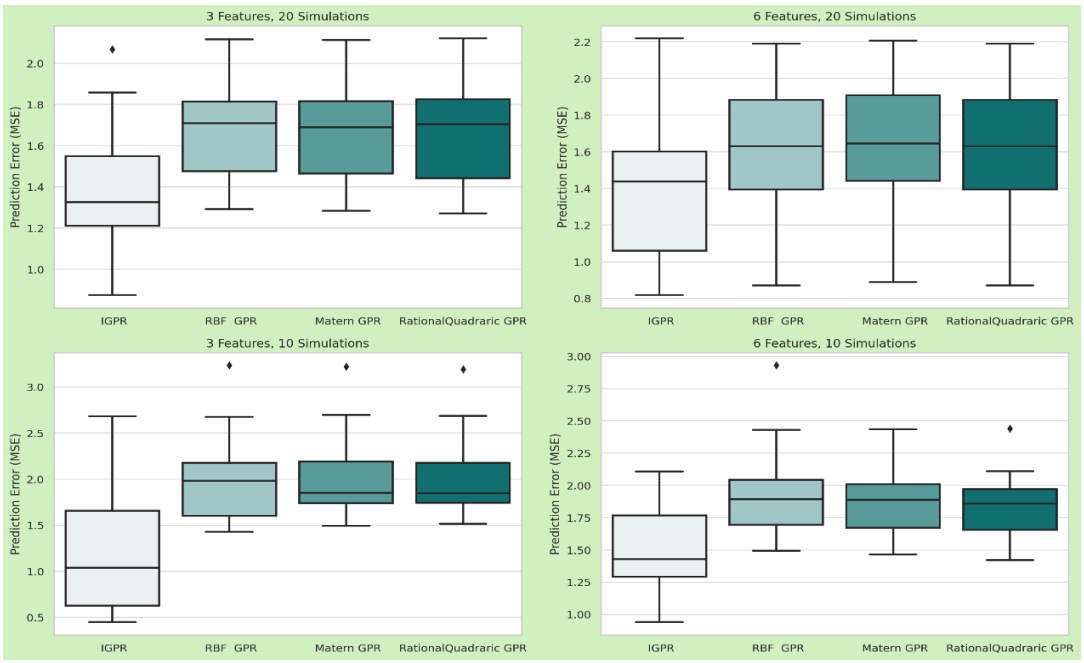

Figure 2: Boxplots for different cases

Table 1: Summary statistics for each model under different experimental conditions

Experiment	IGPR	RBF GPR	Matern GPR	RQ GPR
3 Features 20 Simulations	1.3919	1.6752	1.6667	1.6619
6 Features 20 Simulations	1.3792	1.6084	1.6366	1.6084
3 Features 10 Simulations	1.2032	2.0411	2.0553	2.0532
6 Features 10 Simulations	1.4823	1.9695	1.8766	1.8520

The aforementioned analyses demonstrate that the IGPR model, with its multi-kernel learning and Stacking integration strategy, exhibits robust performance in addressing heterogeneity and high-dimensional features in financial data. In particular, when confronted with high-dimensional features and data with limited simulations, the IGPR model displays superior stability compared to other traditional Gaussian process regression models with a single kernel function.

Table 2 presents a summary of the average root-mean-square error (RMSE) values for the various models across different stocks and a range of experiments. The data presented here provide a foundation for further analyses. As can be observed in Table 2, the mean RMSE values of the IGPR model across all experiments and stocks are typically lower than those of the other models, thereby indicating its notable advantage in the domain of stock price prediction. In particular, the RMSE values of the IGPR model are significantly lower than those of the other models in 10 experiments, indicating that it maintains a high level of prediction accuracy despite the limited sample size.

Table 2: Average RMSE of different models on different stocks

Stock Code	Simulation Count	IGPR	RBF GPR	Matern GPR	RQ GPR
0000019	10	0.8475	1.0063	1.0318	1.0138
0000019	20	0.8999	1.0440	1.0523	1.0408
0000019	40	0.8632	1.0054	1.0244	1.0072
0000521	10	0.8617	0.9758	1.0001	0.9804
0000521	20	0.7678	0.8794	0.9030	0.8843
0000521	40	0.7937	0.9005	0.9302	0.9083
00007278	10	0.5747	0.6890	0.6866	0.6627
0000728	20	0.5261	0.6874	0.6626	0.6495
0000728	40	0.561384	0.6968	0.6907	0.6670

This is a crucial consideration for small-sample prediction tasks in real-world applications. As the number of experiments increases, the root means square error (RMSE) values of all the models decrease, indicating that the models can better learn the changing patterns of stock prices with more experimental data. Nevertheless, the IGPR model exhibits the most pronounced reduction in RMSE and the most consistent performance across diverse stocks, thereby substantiating the adaptability and robustness of the IGPR model in disparate market contexts.

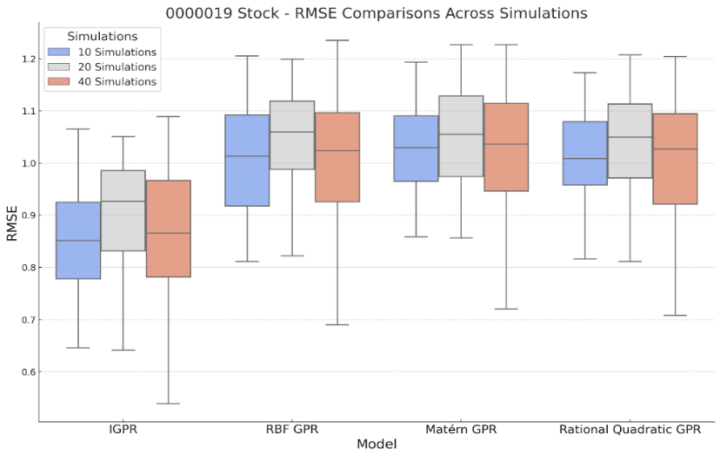

Figure 3: RMSE box plots for stock-0000019 at different number of experiments

As a illustrated, the IGPR model demonstrates superior performance compared to the other models, exhibiting low RMSE values across all stocks, particularly when the sample size is limited. This suggests that the IGPR model is capable of maintaining a high level of predictive accuracy in the context of data scarcity, whereas the other models exhibit a greater disadvantage in this scenario with significantly higher RMSE values. To investigate the precise influence of experiments on the model's performance, Figure 3-5 shows the RMSE trends of the three stocks across varying numbers.

As illustrated by the heatmap, the IGPR model demonstrates superior performance compared to the other models, exhibiting low RMSE values across all stocks, particularly when the sample size is limited. This suggests that the IGPR model is capable of maintaining a high level of predictive accuracy in the context of data scarcity, whereas the other models exhibit a greater disadvantage in this scenario, with significantly higher RMSE values. A detailed examination of the heatmap provides further evidence of the IGPR model's adaptability and its potential for broad applicability in diverse market contexts. To investigate the precise influence of the number of experiments on the model's performance, Figure 6 depicts the RMSE values for three stocks across different numbers of experiments, with color intensity representing the magnitude of the errors. Also Figure 7 depicts the RMSE trends of the three stocks across varying numbers of experiments.

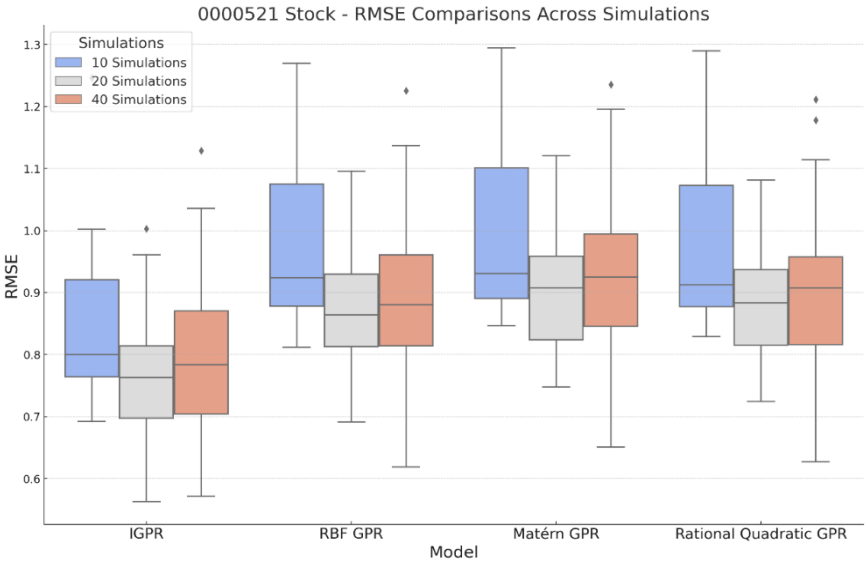

Figure 4: RMSE box plots for stock-0000521 at different number of experiments

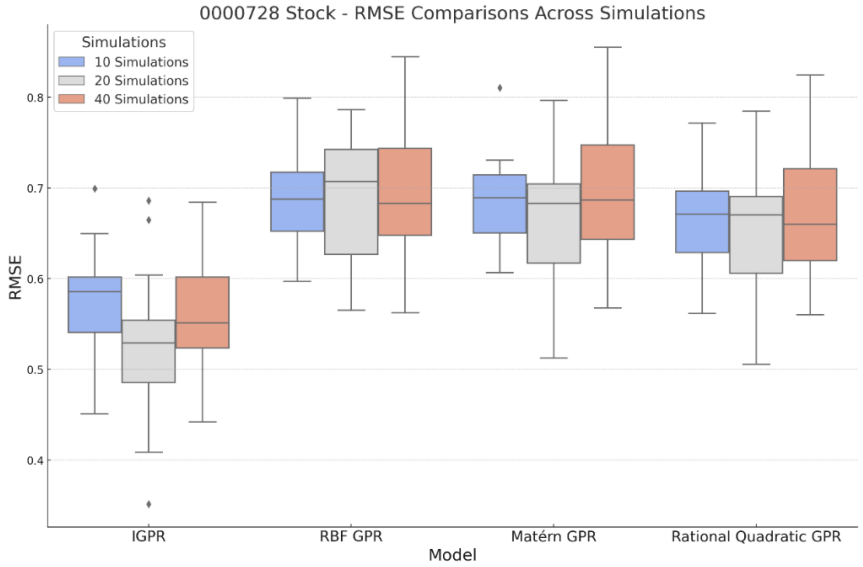

Figure 5: RMSE box plots for stock-0000728 at different number of experiments

As can be seen from the line graph, with the increase of the number of experiments, the RMSE values of all models show a decreasing trend, which indicates that more experimental data can effectively improve the learning ability and prediction accuracy of the models. However, the IGPR model performs particularly well in this process, with the largest decrease in its RMSE value and smaller fluctuations in its RMSE value under different numbers of experiments, especially after 40 experiments, there is almost no significant change in the RMSE value of the IGPR model, which shows a very high prediction stability and market adaptability. In contrast, although the other models also show a downward trend in RMSE values, their changes are small and their performance varies greatly under different numbers of experiments, showing a lack of adaptability under large sample conditions. Finally, in order to comprehensively compare the overall performance of the different models over the 10 experiments, Figure 8 presents a summary of the RMSE boxplots of the three stocks over the 10 experiments. This chart provides a comprehensive comparative view of the predictive ability of different models in small sample situations.

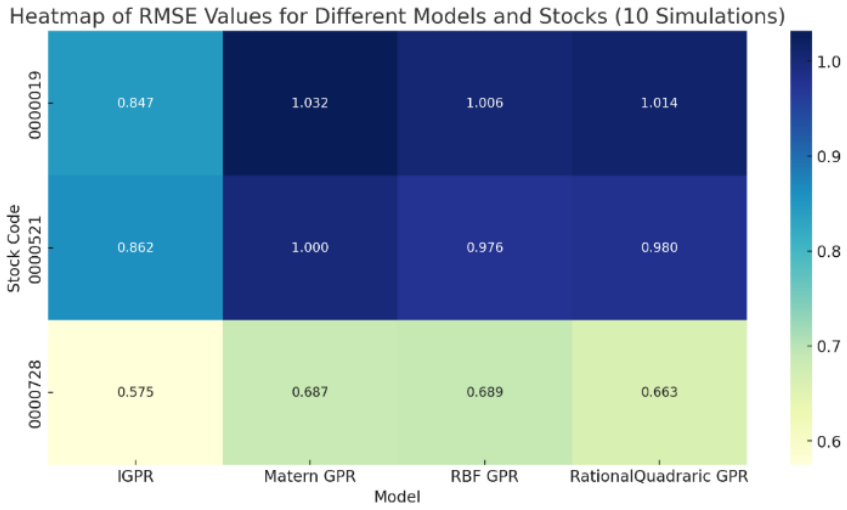

Figure 6: Heatmap of RMSE for different stocks in 10 experiments

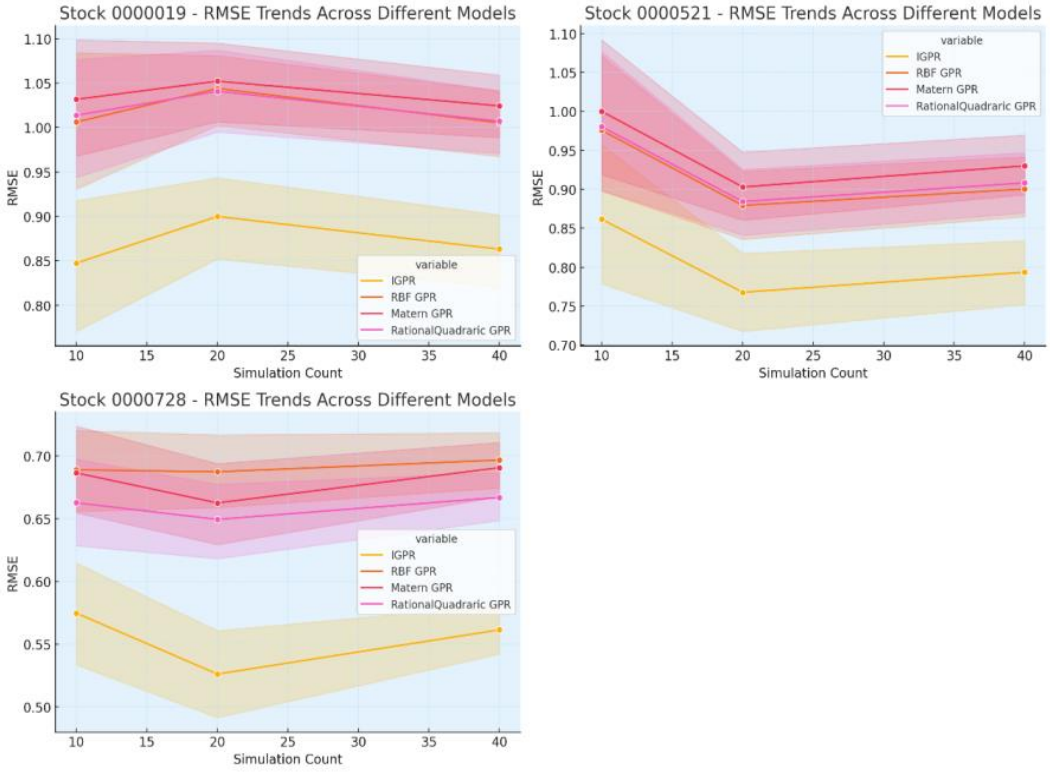

Figure 7: Line graph of RMSE for each stock for different number of experiments

The summary graph clearly demonstrates that the RMSE values of the IGPR model are lower than those of the other models for all stocks. Furthermore, the distribution is more concentrated, indicating that the prediction errors of the IGPR model are smaller and more stable. In contrast, the other models exhibit higher RMSE values on some stocks with a more dispersed distribution, indicating that their performance is not as robust as that of the IGPR model in small samples.

In conclusion, the analysis of the three stocks in detail under different numbers of experiments serves to verify the significant advantages of the IGPR model in stock price prediction. The IGPR model demonstrates consistent efficacy and robustness in financial data prediction, as evidenced by its low RMSE values across both small (10 experiments) and large (40 experiments) sample sizes. Its performance is particularly noteworthy in the context of limited data, where it is able to maintain high prediction accuracy despite the insufficient sample size. This is a challenging aspect to achieve in Gaussian process regression models.

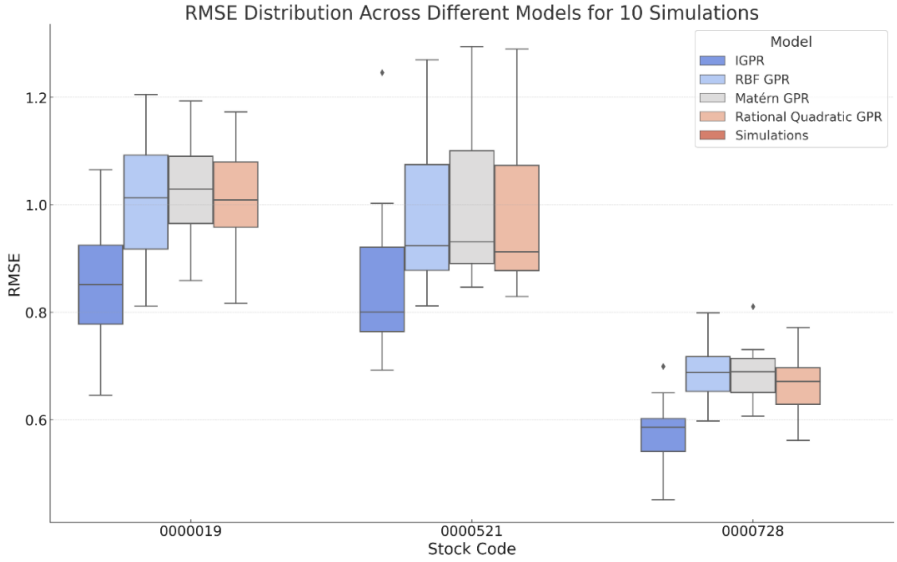

Figure 8: Summary of RMSE box plots for all stocks in the 10 experiments

4. Conclusion and Extension

This study presents the design and validation of a stock price prediction method based on a multikernel Gaussian process regression (IGPR) model, with a particular emphasis on the efficacy of non-Euclidean data within the model. The IGPR model was developed using a multikernel learning strategy, which is capable of effectively integrating the traditional Euclidean factors (e.g., price, turnover) with non-Euclidean factors (e.g., the market turnover rate and the consumer confidence index). The role of these non-Euclidean data in stock price prediction is complex and far-reaching. Traditional regression models often struggle to effectively model these factors. In contrast, the IGPR model, which employs multiple kernels functions, successfully establishes a unified prediction framework in both Euclidean and non-Euclidean spaces, significantly improving the accuracy and stability of the model. By combining simulation experiments with random parameter generation and real data analysis, we conducted a comprehensive and systematic verification of the model's superiority and applicability.

In future research, it would be beneficial to consider that, although the IGPR model performs well in this study, it has a high computational complexity, which may limit its application efficiency in real-time prediction tasks. Further research could seek to enhance the computational efficiency of the IGPR model and investigate its integration with other sophisticated machine learning techniques, such as deep learning and reinforcement learning, with the aim of developing more sophisticated financial market prediction tools. Secondly, future studies could consider a greater variety of stocks and market conditions to comprehensively assess the IGPR model in different market environments.

Furthermore, the successful application of the IGPR model provides a new avenue for research, whereby the role of non-Euclidean factors in financial markets can be further explored. In conclusion, the IGPR model has been demonstrated to possess excellent abilities in stock price prediction. The model has been shown to exhibit significant advantages in particular when dealing with high-dimensional data containing non-Euclidean factors. It is anticipated that the IGPR model will assume a more prominent role in future research in the fields of financial market prediction, risk management and portfolio optimisation, thereby making a significant contribution to the intelligent development of financial markets.

References

[1] Fu, W. Z. Research on stock price prediction based on ARIMA-RF combination model [J]. *Scientific and Technological Innovation*, 2023, (08): 40-43.

[2] Zhou, Z. Y., & He, X. L. Stock price prediction method based on optimized LSTM model [J]. *Statistics and Decision*, 2023, 39(06): 143-148.

[3] Wang, P. D. Stock price analysis and prediction based on multiple linear regression [J]. *Science, Technology and Economic Market*, 2020, (01): 84-85.

[4] Wang, M., & Zhang, M. Stock price prediction and volatility with consideration of high-dimensional macroeconomic information [J]. *Statistics and Decision*, 2022, 38(20): 138-143.

[5] Liu, G. Y., Bao, Y. Y., & Lin, J. G. LASSO-CDRD covariance matrix prediction model based on high-frequency financial data [J]. *Statistical Research*, 2022, 39(09): 145-160.

[6] Xue, H. G., Zhang, R. M., & Hu, C. P., et al. Hedging method based on ridge regression [J]. *Statistics and Decision*, 2012, (05): 77-79.

[7] Lin, M. S., Yang, X. M., & Yang, Z. X. Application of structured maximum margin dual support vector machine in stock prediction [J]. *Computer Engineering and Applications*, 2024, 60(11): 346-355.

[8] Liu, X. L., & Cui, W. H. A classifier for building quantitative strategies based on decision trees [J]. *Pure Mathematics and Applied Mathematics*, 2023, 39(03): 339-349.

[9] Zhou, L. Research on stock multi-factor investment based on random forest model [J]. *Financial Theory and Practice*, 2021, (07): 97-103.

[10] Zhang, X. F., & Wen, X. Research on stock index rise and fall prediction strategy based on XGBoost [J]. *Computer and Digital Engineering*, 2023, 51(03): 686-689.

[11] Zhou, Y. G., Tang, C. W., & Lin, Z. H. Market sentiment, investment factors, and asset pricing: A comparative analysis based on news and social media [J]. *Journal of Econometrics*, 2024, 4(03): 567-587.

[12] Lin, F. J. Research on the relationship between investor sentiment and stock returns: Based on overnight returns [J]. *Investment Research*, 2022, 41(11): 137-159.

[13] Shi, Y. D., Tian, Y. B., & Ma, J. Q., et al. The impact of investor sentiment on cross-sectional stock returns under a multi-factor model [J]. *Investment Research*, 2015, 34(05): 48-65.