Semi-Parametric Functional Kriging Regression Model with L1 Penalty

Rui Chen [1, *], Zhiyong Zhou [2]

[1] College of Mathematics and Statistics, Beihua University, Jilin, China, 132013

[2] College of Mathematics and Statistics, Kashgar University, Kashgar, China, 844000

* Corresponding Author Email: 2784921554@qq.com

Abstract. Partial functional linear models are widely studied and applied models, where the response variable is related to both general random variables and functional random variables. However, with the increasing application of data scenarios involving functional and vector-valued covariates and scalar responses in modern science, this paper proposes a partial functional regression model based on Gaussian processes. On the one hand, the proposed method can flexibly fit the nonlinear connection relationship between the functional covariates and the scalar responses by assuming the existence of a Gaussian process prior between them. On the other hand, for vector-valued covariates, in this paper, while constructing the linear relationship between them and scalar responses, the LASSO regularization technique is used to achieve the purpose of variable selection. Furthermore, in this paper, functional principal component analysis is used as the regularization strategy to approximate the distances between random functions, thereby achieving the approximate calculation of the kernel function matrix. The simulation experiment analysis indicates that the proposed method has higher prediction accuracy compared with the benchmark model and can effectively identify irrelevant variables. The actual data analysis also confirmed the comprehensive performance of the proposed method.

Keywords: Functional data regression, Variable selection, Gaussian process, L1 regularization.

1. Introduction

With the continuous advancement of data collection and acquisition technologies, the amount of high-frequency and densely recorded data is increasing day by day. This type of data is called functional data and has wide applications in multiple scientific fields, such as biomedicine, economic management, and computer science. Unlike ordinary high-dimensional data, functional data not only has high-dimensional characteristics but also has a high correlation among time series. If it is only regarded as high-dimensional data, it is effortless to ignore its inherent functional characteristics. Therefore, when the explanatory variable is functional data, the classical regression model and the high-dimensional regression model will no longer be applicable. For example, Hyndman and others created a system for predicting functional time series by using functional principal component regression and functional partial least squares regression, and they showed that this method works well with data on population death and birth rates [1]. Mollenhauer et al. explored the problem of optimal linear prediction in infinite-dimensional Hilbert spaces and proposed an asymptotic optimal estimation method [2]. Yao et al. proposed an adaptive basis function layer deep learning framework. By optimizing the basis functions through end-to-end learning, the prediction accuracy of functional data classification and regression was significantly improved [3].

Researchers have carried out extensive explorations around the processing and modeling of functional data, gradually forming a mature methodological system: Firstly, when working with individual pieces of functional data, Fourier basis expansion breaks it down into different frequencies using sine and cosine functions, B-spline basis expansion creates smooth curves with small polynomial sections, and wavelet transform finds local features by analyzing the data at different levels of detail. These smoothing processing techniques effectively eliminate the noise interference in the original data. Wan Anis Farhah and Wan Amir et al. proposed a flexible smoothing method based on β splines and combined roughness penalty and generalized cross-validation (GCV) to

optimize the smoothing parameters, enhancing the flexibility of the model and making it suitable for processing complex functional data [4]. For example, Yao Y proposed a deep learning framework for end-to-end learning, which adaptively optimizes the basis parameters of B-splines and combines wavelet transform to extract multi-scale features [5]. Then, functional principal component analysis can pull out important information from the smoothed random function, convert the infinite-dimensional function space into a simpler, finite-dimensional score, and set the stage for further statistical analysis. For instance, Zhang J. et al. proposed a robust FPCA method based on Kendall's tau function, which solved the estimation bias problem of traditional FPCA in non-Gaussian numbers [6]. Huang Y et al. proposed the parametric FPCA method, using polynomial functions to approximate the functional principal components (FPCs), significantly improving the interpretability of FPCs [7]. The functional linear model has become a classic tool for analyzing the relationship between functional data and response variables, as it naturally extends multiple linear regression in the function space and offers a simple mathematical structure and interpretability. For example, Yao Y combines FLM with CNN, optimizes the base parameters of B-splines through end-to-end learning, improves the classification accuracy in speech emotion recognition, and demonstrates the potential of FLM in dynamic data modeling [5]. Chen D et al. proposed the Bayesian FLM framework, which processes uncertainty through adaptive basis function selection and solves the problems of poor interpretability and prediction accuracy of this model [8].

In practical application scenarios, data structures often exhibit mixed characteristics, and it is not uncommon for functional covariates and Euclidean covariates (such as vector-based data like age and gender) to coexist. Ignoring the information of the Euclidean covariates may result in insufficient model fitting ability. In view of this, some scholars have proposed partial functional linear models. For instance, Zhang Xinyu et al. proposed the optimal weight selection criterion for the model averaging method under the partial number linear model and presented the asymptotic optimality theory of the model average estimator [9]. Wang Huiwen et al. extended the functional linear model to the generalized linear framework, dealt with discrete and attribute response variables, and combined B-spline basis functions and functional principal component analysis to enhance the flexibility of the model [10]. However, the data relationships in the real world are far more complex than theoretical assumptions. In the field of medical and health care, the influence of functional predictor variables on response variables often shows highly nonlinear characteristics between the vital sign curve of patients and the severity of diseases and between the curve of asset price fluctuations and the rate of return in the financial market. The traditional functional linear model, due to its strict linear assumption, has obvious limitations when dealing with such complex relationships and is difficult to accurately depict the underlying patterns behind the data. Furthermore, simply incorporating all the Euclidean covariates will not only increase the computational burden but also may introduce redundant information and reduce the generalization performance of the model. More importantly, when we don't know how the variables are related, finding and selecting the key factors that really influence the outcome is essential for making the model easier to understand and improving its prediction accuracy.

In response to the above challenges, this paper proposes a functional Gaussian process regression model based on LASSO. This model adopts a dual-path modeling strategy: The model applies a linear assumption to the relationship between vectoring value variables and responses. With the help of the sparse constraint characteristics of LASSO, the automatic selection of Euclidean covariates is achieved, effectively eliminating redundant variables. On the other hand, by introducing the Gaussian process prior to describe the relationship between functional covariates and responses, the flexibility of the Gaussian process in nonlinear modeling is fully utilized to break through the expression bottleneck of traditional linear models. To make calculations easier, a new method called functional principal component truncation is used to estimate the distance between functions, which greatly lowers the complexity of the kernel function matrix and speeds up the model training process. Finally, a new framework is created to optimize parameters based on maximizing the negative logarithmic likelihood function of the LASSO penalty, ensuring that the model fits well while also selecting the

right variables. By conducting organized simulation tests and analyzing real data, we have confirmed that this model has clear benefits in predicting accuracy and identifying variables, offering a new and effective way to model complex relationships in functional data.

2. Methodology

The existing data are $X(t),\ Z_i,\ Y_i,\ i = 1,2,\cdots n$. Among them, $X(t)$is the functional data related to time, Z_i is an i-dimensional vector-valued data vector. and Y_i is the response variable data.

Let's assume

$$Y_i = z_i^T\beta + f\big(X(t)\big) + \varepsilon \tag{1}$$

The relationship between $X(t)$ and Y_i is unknown. The vector-valued data Z_i shows a linear relationship with Y_i, but the exact nature of this relationship is uncertain. β is the parameter vector of dimension i. used to select vector-valued data that is uncorrelated with Y_i,ε represents the noise factor, $\varepsilon \sim (0, \sigma_n^2)$. Assuming $f \sim GP(0, k)$, then Y_i will also follow a Gaussian process, that is, $Y_i \sim (z^T\beta, k)$. Here, k represents the matrix of the Gaussian kernel function.

$$K\big(x_1(t), x_2(t)\big) = \exp\left[-\frac{\|x_1(t)-x_2(t)\|^2}{2\sigma^2}\right] \tag{2}$$

The σ^2 is parameter in the Gaussian kernel matrix. Solving the problem of representing $X(t)$ is something we must take into consideration. Given that the function is time-related functional data, we first smooth $X(t)$ by using sine and cosine functions from the Fourier basis to ensure the stability and interpretability of the overall results. Then, we obtain its approximate representation through functional principal component analysis. Perform the basis expansion on $x_1(t)$ and $x_2(t)$, and obtain this formula

$$x_1(t) = a^T \cdot \phi, x_2(t) = a_2^T \cdot \phi \tag{3}$$

ϕ is a basis function, so $\|x_1(t) - x_2(t)\|^2$ can be approximated as the inner product of their functional variables, that is

$$\begin{aligned}\|x_1(t) - x_2(t)\|^2 &= \langle x_1(t) - x_2(t), x_1(t) - x_2(t)\rangle \\ &= (a_1 - a_2)^T < \phi, \phi > (a_1 - a_2) \\ &= \|a_1 - a_2\|^2\end{aligned} \tag{4}$$

So the Gaussian kernel matrix is approximately expressed as

$$k\big(x_1(t), x_2(t)\big) = \exp\left[-\frac{\|a_1-a_2\|^2}{2\sigma^2}\right] \tag{5}$$

The construction of the basis and covariance functions leads to the formulation of the likelihood function for the response variable Y, which can be expressed as

$$L(y_1, y_2, \cdots y_n, z^T\beta, k) = \Pi_{i=1}^n \frac{1}{\sqrt{2\pi k^2}} \exp\left[-\frac{(y_i - z^T\beta)^2}{2k^2}\right] \tag{6}$$

To facilitate subsequent calculations, we will estimate the parameters using the negative logarithmic minimum likelihood function. Since the relationship between Z_i and Y_i is unknown, we adopt LASSO regression. We apply $L1$ penalty to β after the likelihood function to select variables for Z_i. The process is expressed as

$$\text{argmin}_\beta - \ln L + \lambda\|\beta\|_1 \tag{7}$$

The penalty coefficient λ is set to 0.1. Finally, we use the Newton iterative algorithm to estimate the unknown values of $\hat{\beta}$ and $\hat{\sigma}^2$ in the above equation. Since there are n samples, a joint normal distribution of $n + 1$ dimensions can be obtained through the Gaussian process. Therefore, we can

update Y_{n+1} based on the information of $Y_1, Y_2, \cdots Y_n$, and Y_{n+1} also follows a Gaussian process. The Bayes formula yields the following deduction:

$$y^*|y_1, y_2 \cdots y_n \sim N(\mu^*, \Sigma^*) \tag{8}$$

$$\mu^* = (z^*)^T \hat{\beta} + k^* k^{-1} \left(y - z^T \hat{\beta}\right) \tag{9}$$

$$\Sigma^* = k^{**} - (k^*)^T k^{-1} k^* \tag{10}$$

$k^*$is an n*1 matrix, and k^{-1} is an n*n matrix. According to the maximum posterior probability criterion, the point with the highest probability in the normal distribution is the mean point. Therefore, we can use the mean as the estimated value, that is

$$\hat{y}^* = \mu^* \tag{11}$$

Next, we will use a flowchart to visually represent our method, as shown in Figure 1.

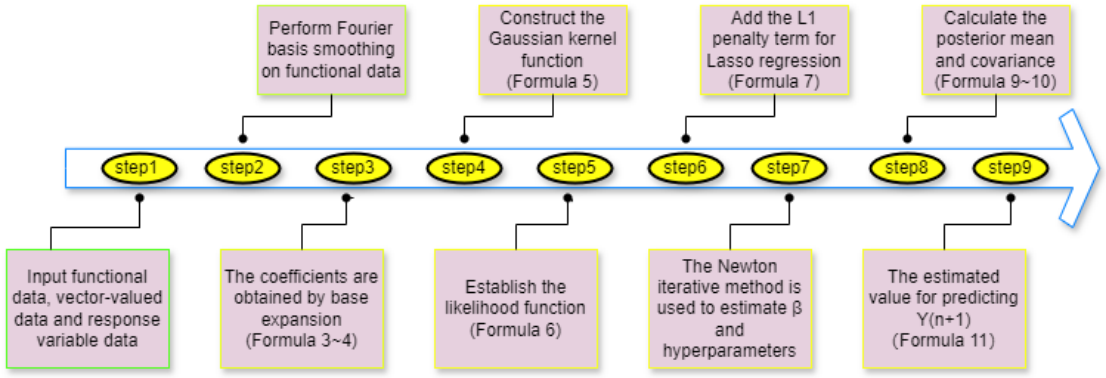

Figure 1 Method flowchart

3. Data analysis

3.1. Simulation experiment

Consider the following partial functional regression model

$$Y_i = z_{ij}^T \beta_j + f\left(x_i(t)\right) + \varepsilon, i = 1,2, \ldots, 100 \tag{12}$$

Among them, z_{ij} follows a standard multivariate normal distribution, with its mean being a zero vector and the covariance matrix Σ being the identity matrix. To identify which vector values are related to data z_{ij} and Y_i, we set β to have zero values for variable selection. We also consider that the dimension of β ranges from 5 to 6, and the number of zeros in β increases with the increase in dimension, from 2 to 3. The type of f is set to be either $\sin$ or $\cos$ functions, totaling 4 different settings. Random error term $\varepsilon \sim (0, \sigma_n^2)$. We use Fourier basis functions to generate functional data and coefficient functions. Subsequently, we perform the functional principal component basis expansion, extract and retain the basis expansion coefficients ranging from 5 to 6. The generation of functional data and coefficient functions can be respectively expressed as

$$x_i(t) = \sum_{j=1}^{100} a_{ij} \left(\frac{1}{k}\right)^{\frac{3}{2}} \phi_j(t) \tag{13}$$

$$\beta(t) = \sum_{j=1}^{100} \left(\frac{1}{k}\right)^2 \phi_j(t) \tag{14}$$

Among them, a_{ij} is a random term following a normal distribution, which will randomly select z_{ij} that follows a normal distribution. $\phi_j(t)$ is a Fourier basis function, $t \in [0.01,1]$ and the step

size being 0.01. Thus, $\langle x_i(t), \beta(t) \rangle$ can be calculated, and subsequently, $f(x_i(t))$ can be simulated, which are respectively

$$f_1(x_i(t)) = 0.01 \times \sin(\langle x(t), \beta(t) \rangle) \tag{15}$$

$$f_2(x_i(t)) = 0.01 \times \cos(\langle x(t), \beta(t) \rangle) \tag{16}$$

In order to measure the prediction accuracy of the model's response variable and the model's volatility, we conducted 10 Monte Carlo experiments. We used 40 samples as the test set and 60 samples as the training set, and selected MSE, MAE, and MRE as the evaluation indicators. The details are as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{100} (y_i - \hat{y}_i)^2 \tag{17}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{100} |y_i - \hat{y}_i| \tag{18}$$

$$\text{MRE} = \frac{1}{n} \sum_{i=1}^{100} \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{19}$$

The smaller the index, the higher the accuracy of the response variable. When the estimated β is more accurate and the evaluation index value is lower, it indicates that our model is better.

To save space, we only provide the results for 6 dimensions. The estimated simulation results are shown in Figures 2 and 3.

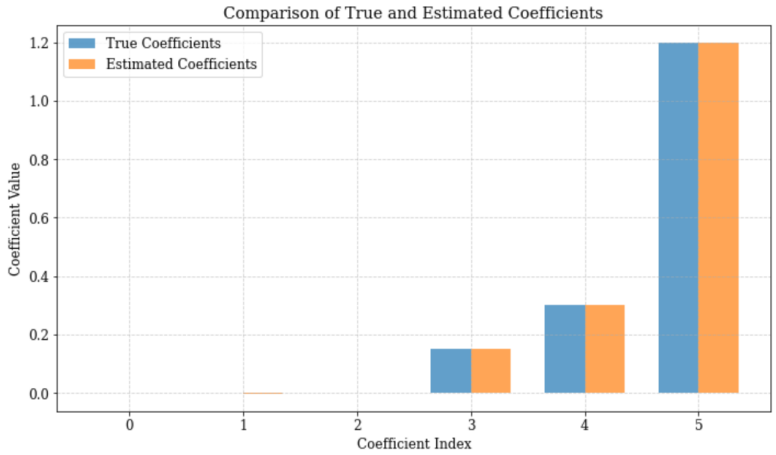

Figure 2 Comparison of the true value and the estimated value of β when the function is equal to $f_1(x_i(t))$

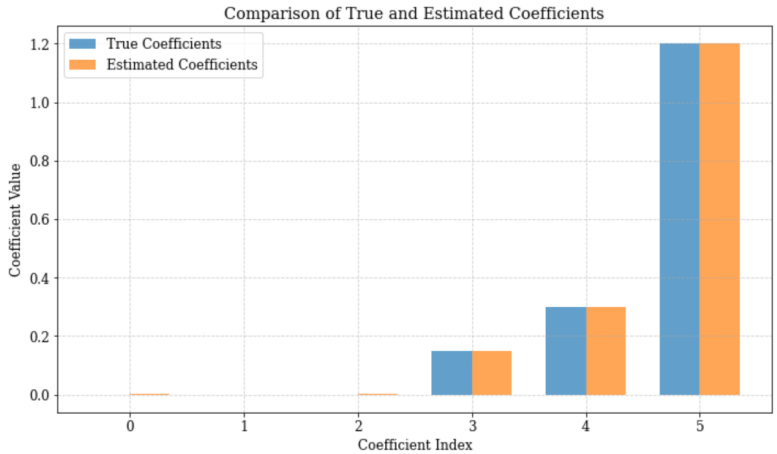

Figure 3 Comparison of the true value and the estimated value of β when the function is equal to $f_2(x_i(t))$

It is evident that the β values estimated by our model are all of high accuracy. To verify whether the coefficients of the β values estimated as the variables for selection are accurate, we also conducted a comparison of the test set errors, as shown in Figures 4 and 5.

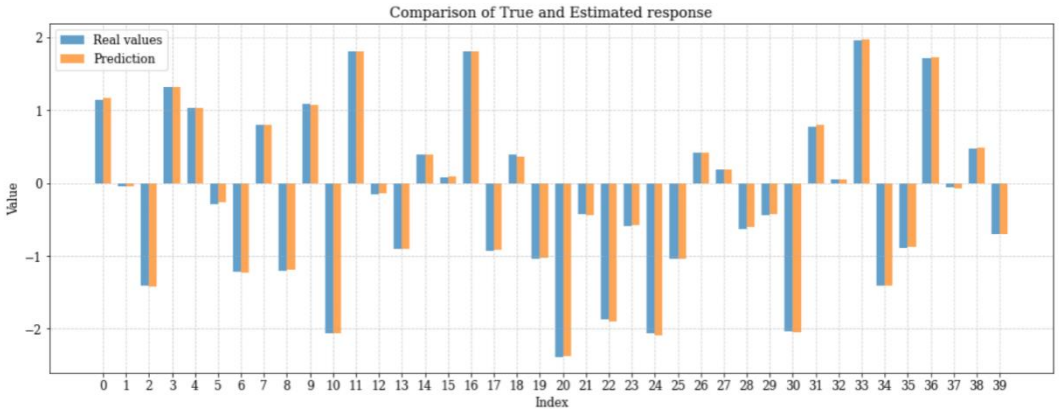

Figure 4 Error analysis of β in the test set when the function is equal to $f_1(x_i(t))$

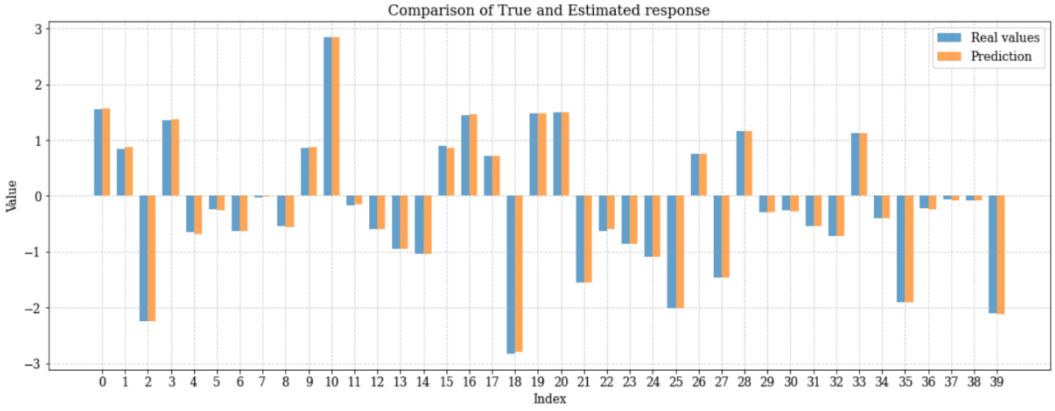

Figure 5 Error analysis of β in the test set when the function is equal to $f_2(x_i(t))$

It is evident that both the error of the test set and the prediction are relatively small. Therefore, the β value we estimated can be used for variable selection in this model without any problem. In addition, in the Monte Carlo simulation experiment, we considered the comparison of different models, namely XGBoost, SVR with the base and Gaussian kernel functions, elastic net, and Gaussian regression process model. As depicted in Figures 6 and 7 and Tables 1 and 2.

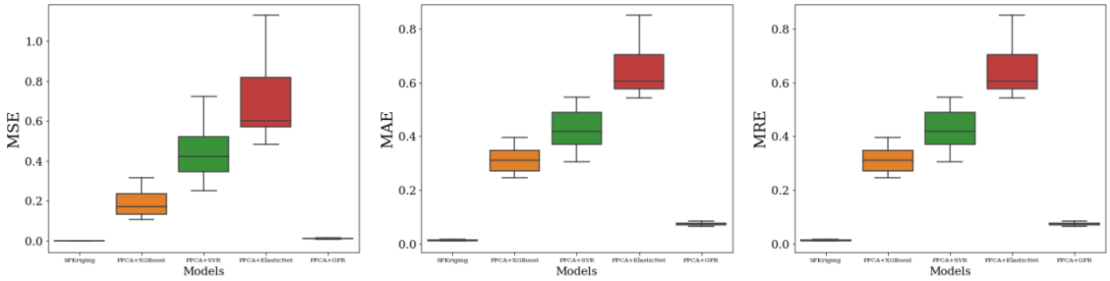

Figure 6 Analysis of box plots for each model when the function is $f_1(x_i(t))$

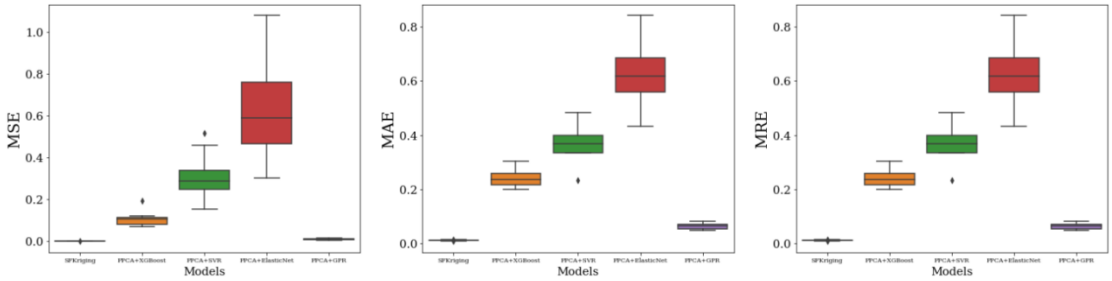

Figure 7 Analysis of box plots for each model when the function is $f_2(x_i(t))$

Table. 1 Data analysis of evaluation indicators for each model when the function is equal to $f_1(x_i(t))$

	Mean	SE(MSE)	AE	SE(MAE)	AE	SE(MRE)
SFKriging	0.0002	0.0000	0.0129	0.0013	0.0690	0.0404
FPCA+XGBoost	0.1880	0.0715	0.3136	0.0496	0.8447	0.4147
FPCA+SVR	0.4513	0.1604	0.4215	0.0814	0.5324	0.0919
FPCA+ElasticNet	0.7166	0.2352	0.6501	0.1090	0.9259	0.3066
FPCA+GPR	0.0110	0.0021	0.0736	0.0062	0.1945	0.0562

Table. 2 Data analysis of evaluation indicators for each model when the function is equal to $f_2(x_i(t))$

	Mean	SE(MSE)	AE	SE(MAE)	AE	SE(MRE)
SFKriging	0.0002	0.0001	0.0121	0.0018	0.1477	0.0845
FPCA+XGBoost	0.1051	0.0352	0.2415	0.0323	3.6300	2.4700
FPCA+SVR	0.3117	0.1073	0.3706	0.0711	1.5832	0.8718
FPCA+ElasticNet	0.6221	0.2267	0.6255	0.1159	3.0049	1.5321
FPCA+GPR	0.0085	0.0034	0.0636	0.0119	0.6878	0.4785

The box plot shown in the figure clearly indicates that the interquartile range of SFKriging is significantly lower than that of other models, which demonstrates that our model is sufficiently robust. Moreover, most of the MSE, MAE, and MRE values are close to 0 and the standard deviation is extremely small, reflecting that our model's predictions are highly stable and the overall error is relatively low. FPCA + GPR is suboptimal, but its MRE is significantly higher than MAE. This might be because the kernel function of GPR does not match, resulting in poor function fitting for high-frequency or small-scale changes, making this model more sensitive to outliers. FPCA + ElasticNet performs the worst, with the highest MSE and MAE values and a large standard deviation. This is because this model assumes a linear relationship between the response variable and the principal components, thus unable to capture nonlinear relationships, leading to systematic bias and low model stability. It may also be due to insufficient tuning of the regularization parameters, and the performance of ElasticNet is highly dependent on hyperparameters. Therefore, the final result is that it has the poorest performance. Analyzing the experimental results, it is easy to see that our model has achieved the best results. Both the model's stability and the prediction accuracy are significantly higher than those of other models.

3.2. Real data analysis

In this section, we apply the proposed functional semi-parametric Kriging regression model to the spectral data set of meat samples measured by the near-infrared spectrometer to analyze the performance of the model in practical problems. The data background of this article will predict the protein content of the meat samples based on the curve data of the spectrometer for the meat samples

and the content of fat and moisture.The data set of the meat samples was obtained from http://lib.stat.cmu.edu. This data set has 240 samples, each of which contains the moisture, fat, and protein content of the meat, and includes a total of 100 absorption spectrum data measured by the near-infrared spectrometer in the meat samples. Using MSE and MAE as the evaluation indicators of the model, in addition to the comparison of the 4 models in the simulation experiment, in this experiment, PCA was also integrated with the 4 models to be compared with our model. With β set to 6 and the function as $f_2(x_i(t))$, The experimental results are shown in Figure 8 and Tables 3 and 4.

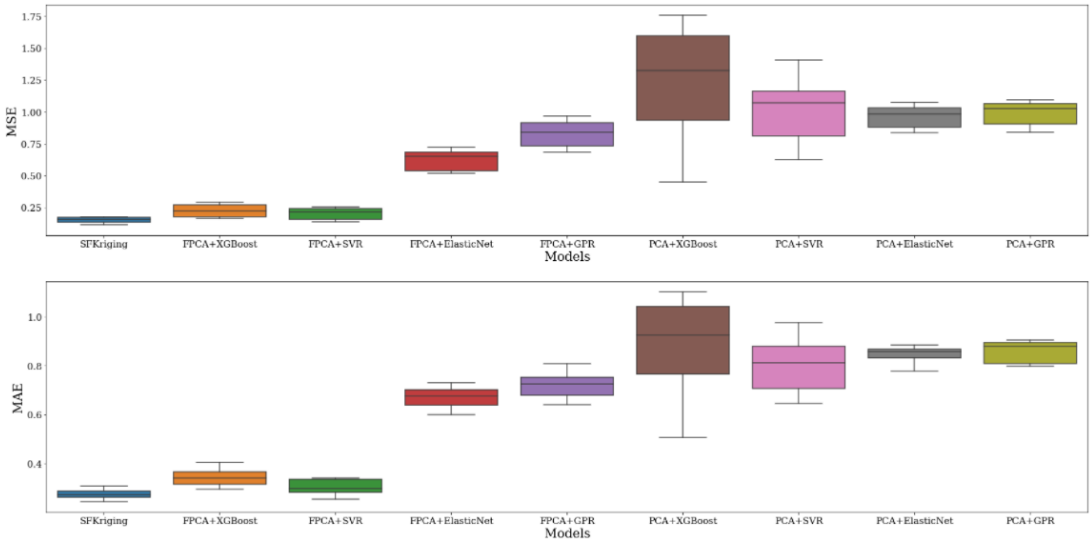

Figure 8 Box plot analysis of the FPCA series models and the PCA series models

Table.3. Analysis of Evaluation Indicators for FPCA Series Models

	SFKriging	FPCA+XGBoost	FPCA+SVR	FPCA+ElasticNet	FPCA+GPR
Mean	0.1525	0.2233	0.2037	0.6231	0.8300
SE(MSE)	0.0240	0.0495	0.0462	0.0819	0.1052
AE	0.2755	0.3449	0.3043	0.6682	0.7221
SE(MAE)	0.0194	0.0382	0.0326	0.0431	0.0521

Table.4. Analysis of Evaluation Indicators for PCA Series Models

	PCA+XGBoost	PCA+SVR	PCA+ElasticNet	PCA+GPR
Mean	1.2080	1.0329	0.9625	0.9860
SE(MSE)	0.4753	0.2632	0.0859	0.1017
AE	0.8657	0.8017	0.8446	0.8659
SE(MAE)	0.2194	0.1135	0.0358	0.0459

From the chart, it can be seen that our model SFKriging performs the best, with MSE and MAE significantly lower than those of other models; FPCA + GPR performs the worst among the FPCA series models, but still has certain advantages compared to the PCA series models; the overall performance of the FPCA series models is better than that of the PCA series models. In terms of stability, the standard deviation of SFKriging is the lowest, indicating that the prediction results of

our model are the most stable, while the standard deviation of PCA + XGBoost is the highest, suggesting that the performance of this model fluctuates significantly. Overall, SFKriging is the model with the best comprehensive performance, with the best accuracy and stability; the FPCA method significantly improves the model's performance, especially in XGBoost and GPR; the overall performance of the PCA series models is poor, which may be due to the continuous nature and shape-dependent characteristics of the data, causing PCA to be unable to utilize the functional structure and thus perform poorly.

4. Conclusion and Outlook

The partial functional Gaussian process regression model based on LASSO is a highly flexible model that encompasses the classical multivariate linear model, functional linear model, Gaussian process regression model, and LASSO regression model, thereby inheriting the advantages of these models. The proposed method in this paper assumes a Gaussian process prior between the functional covariates and the scalar response to fit their nonlinear connection. Subsequently, while constructing the linear relationship between the vector-valued covariates and the scalar response, the LASSO regularization technique is employed to achieve variable selection for the vector-valued covariates. Finally, the distance between random functions is approximated through functional principal component analysis, thereby obtaining the approximate calculation of the kernel function matrix. We conducted numerous simulation experiments and real data analysis of meat samples, and the results all indicated that the proposed method in this paper has higher prediction accuracy and stability compared to other benchmark models.

For the model proposed in this paper, further extensions are certainly possible. On one hand, the choice of truncation number could be determined by an ensemble learning model, that is, by assigning a larger number of recursive stages in the functional principal component analysis and constructing an ensemble learning model based on this sequence. On the other hand, considering the nonlinear relationship of Euclidean covariates and the interpretability of variables, it is also possible to use additive models for processing.

References

[1] Hyndman, R. J., & Shang, H. Functional time series forecasting. Journal of the American Statistical Association, 2009,104(488), 1542-1553.

[2] Mollenhauer, M., et al. Optimal linear prediction with functional observations. arXiv preprint arXiv,2023,2401.06326.

[3] Yao, F., et al. Deep Learning for Functional Data Analysis with Adaptive Basis Layers. Proceedings of the 39th International Conference on Machine Learning.2022.

[4] Wan Anis Farhah Wan Amir.Flexible functional data smoothing and optimization using beta spline, Mathematics,2024,10.3934/math.20241126

[5] Yao Y et al. Adaptive basis function layer framework for functional data analysis[J]. Acta Automatica Sinica, 2022, 48(10): 2234-2245.

[6] Zhang J, Zhong R. Robust functional principal component analysis for non-Gaussian longitudinal data[J]. Journal of Multivariate Analysis, 2021, 188: 104770.

[7] Huang Y et al. Parametric functional principal component analysis [J]. Biostatistics，2017, 73 (3): 802-812.

[8] Chen D, Liu C. Bayesian functional linear models with adaptive basis selection[J]. Journal of the Royal Statistical Society: Series B, 2023, 85(2): 451-478.

[9] Zhang Xinyu, Zhu Rong, Zou Guohua. Model Averaging Methods for Partially Linear Models.J.Sys.Sci.&Math.Scis.38(7),2018,777-800

[10] Wang Huiwen, Huang Lele, Wang Siyang. Generalized Linear Regression Model Based on Functional Data. Journal of Beijing University of Aeronautics and Astronautics.10.13700/j.bh.1001-5965.2015.0078.