Research on Partial Functional Nonparametric Regression Models Based on Gaussian Process Prior Modeling

Xiaohan Liu *

School of Science, XI'AN POLYTECHNIC UNIVERSITY, Xi'an, China, 710600

* Corresponding Author Email: 19391864268@163.com

Abstract. Aiming at data scenarios simultaneously involving vector - type and functional - type covariates, this paper proposes a partial functional regression model grounded in a Gaussian process prior. The core innovation of the method resides in: introducing a Gaussian process prior to the association structure between the functional - type covariate and the scalar response variable, while assuming the connection function between the vector - type covariate and the scalar response variable exists within the reproducing kernel Hilbert space. This dual specification empowers the model to effectively capture the nonlinear relationship between the vector - type covariate and the response variable via flexible adjustment of the kernel function, and also accounts for the characteristic depiction of the functional - type covariate. Results of actual data analysis validate the superiority of the proposed method, with its predictive performance significantly outperforming existing benchmark approaches, thereby furnishing a more effective solution for regression modeling in complex covariate scenarios.

Keywords: Complex Data Analysis, Semi - parametric Regression, Gaussian Process, Kernel Method.

1. Introduction

In modern data analysis, scenarios involving mixed data that simultaneously contain vector - valued data and functional data have become increasingly prevalent. The modeling of such data is of great significance for revealing the relationships among variables and enhancing prediction accuracy. For example, Wang Hongjian, et al. [1], in their exploration of the application and development of Bayesian priors in complex data modeling, systematically summarized the applicable scenarios and optimization directions of Gaussian process priors, thereby providing support for the theoretical depth of models. Specifically, in medical research, Liu Hong, Chen Yang, Zhao Lei, et al. [2], analyzed how patients' vector - type physiological indicators and ECGs affect cardiovascular disease risk, Zhao Gang , et al. [3] , Mixed data modeling is of great significance. Ignoring data characteristics easily leads to model bias, so effectively conducting this modeling is a key issue to solve.

For the modeling challenges involving both vector - valued and functional data, traditional regression approaches like random forests and neural networks, while capable of handling nonlinear relationships, frequently treat functional data as discrete point sets or vectors after simplistic dimensionality reduction. This inherently overlooks their intrinsic functional traits (e.g., continuity, smoothness, dynamic trends), making it arduous for models to capture the fundamental patterns of the data and ultimately undermining regression or prediction performance. To address this, scholars have put forward the Partial Functional Linear Model (PFLM). This model posits a linear relationship between the response variable and vector - valued covariates, while establishing an associative link with functional covariates via linear function mapping, thus better preserving the characteristics of functional data. For instance, Huang Shanshan and Yang Jingping [4] provided a review of functional data analysis methods, systematically expounding the basic framework of functional data analysis and laying a theoretical groundwork for the advancement of PFLM. Yao Fang, et al. [5] , building on the partial functional linear model, investigated the relationship between the air quality index, meteorological factors (vector - valued), and pollutant concentration curves (functional), validating the model's efficacy in mixed - data modeling. Zhang Ning and Wang Yan [6] enhanced the estimation method of PFLM, boosting the model's adaptability to high - dimensional vector

covariates.However, in reality, the relationship between the response variable and covariates (notably vector - valued covariates) is often far from strictly linear and may entail intricate nonlinear associations. In such circumstances, the linear assumption of PFLM becomes difficult to satisfy, restricting the model's applicability. Lin Hao , et al. [7] explored parameter optimization and application of the generalized additive model in the collaborative modeling of mixed data. For mixed scenarios involving vector - valued data (e.g., population and economic indicators) and functional data (e.g., consumption curves and environmental monitoring time series), they extended the Generalized Additive Model (GAM) framework. By leveraging spline basis functions to flexibly depict the continuity and dynamic trends of functional data and integrating the fitting of linear/nonlinear effects of vector - valued variables, in empirical studies such as economic growth prediction (integrating GDP vectors and consumption - curve functional data) and environmental quality assessment (integrating pollutant - concentration vectors and meteorological - time - series functional data), they verified the model's capacity to preserve the characteristics of mixed data. Compared with the traditional PFLM, it is more adept at accommodating complex nonlinear associations, offering a general methodology for the collaborative modeling of multi - type data. Zhao Meng ,et al. [8] , focusing on economic prediction scenarios, integrated vector - valued indicators such as GDP and unemployment rate with functional data like consumption and investment time - series curves. They refined PFLM within the Bayesian framework, introducing dynamic prior distributions to characterize the time - varying associations of economic variables. In the empirical study of economic growth prediction in OECD countries, it showcased the ability to accurately depict the structure of mixed data, furnishing a novel approach for optimizing economic prediction models.

To tackle nonlinear associations, extant research predominantly employs two categories of methods. First, methods rooted in basis function expansion (e.g., spline functions, Fourier bases) convert nonlinear relationships into linear forms. They capture nonlinear characteristics by expanding the representation space of functional covariates (Sun Yue, et al. [9]). Second, kernel methods are introduced. These leverage the nonlinear mapping capability of deep kernel learning to establish nonlinear links between response variables and covariates (Gao Xiang, et al.[10]).However, these methods either suffer from the subjectivity inherent in basis function selection or confront computational challenges posed by high - dimensional kernel matrices, and related research is still striving for breakthroughs. For instance, Chen Xi and Huang Wei [11], in the modeling of agricultural mixed data, integrated multi - kernel learning with adaptive basis function expansion and verified the adaptability of the improved method to nonlinear associations. Nonetheless, the generalizability of the method in complex scenarios remains to be further explored.Thus, for the extensive nonlinear relationships potentially present in mixed data, developing a more flexible and adaptable modeling approach holds significant theoretical and practical value. Grounded in the above research context, this paper proposes a nonparametric regression model applicable to mixed data comprising functional and vector - valued types.

Its predictive performance not only surpasses that of traditional linear models but also outperforms existing nonlinear modeling methods. As such, it provides an effective solution to regression problems in the context of mixed data.

2. Theories and Methods

We propose a new model as follows.

$$y = f(x(t)) + g(z) + \varepsilon \tag{1}$$

where, y denotes the response variable; $X \in H$ represents a functional predictor, with H being a Reproducing Kernel Hilbert Space (RKHS); z is a vector - type predictor; β stands for the coefficient vector of vector - type variables; and ε is a random error term, satisfying the assumption of a mean value of 0.

Suppose the observed data are$\{y_i, z_i, x_i(t)\}_{i=1}^n$, From the nonparametric regression model with mixed data of the following functional type and vector-valued type:

$$x(t) \in H = \{\sum_i \quad a_i \Phi_i(t) | a_i \in \mathbb{R}\} \tag{2}$$

$$y = f(x + \mathcal{H}, z) + \varepsilon == f(x(t)) + g(z) + \varepsilon z^\top \beta + f(x(t)) + \varepsilon \tag{3}$$

The Gaussian process characterizes the correlation between functions through a kernel function. Define the Gaussian kernel function:

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2l^2}\right) \tag{4}$$

Construct an $n \times n$ kernel matrix K, where the element $K_{pq} = k(\xi_p, \xi_q)$, that is:

$$\mathbf{K} = \begin{bmatrix} k(\boldsymbol{\xi}_1, \boldsymbol{\xi}_1) & \dots & k(\boldsymbol{\xi}_1, \boldsymbol{\xi}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{\xi}_n, \boldsymbol{\xi}_1) & \dots & k(\boldsymbol{\xi}_n, \boldsymbol{\xi}_n) \end{bmatrix} \tag{5}$$

Here,l is the length - scale parameter, which controls the smoothness of the kernel function. The kernel function induces a Reproducing Kernel Hilbert Space (RKHS), and its norm is defined as:

$$\| f - f_h \|_{\mathcal{H}}^2 = \int_{\mathcal{X}} (f(x) - f_h(x))^2 \rho(x) dx \tag{6}$$

Suppose the observation value y is jointly generated by a linear term, a Gaussian process, and noise, that is:

$$y = z^T \beta + f(x) + \epsilon \tag{7}$$

Here,$z^T\beta$ is a linear combination of kernel functions (z is the feature vector, β is the coefficient), is a $\text{zero} - \text{mean Gaussian process}$, and $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ is independent noise. At this time, the distribution of y can be expressed as:

$$y \sim \mathcal{GP}(z^T \beta, k + \sigma_n^2 \delta(\cdot, \cdot)) \tag{8}$$

Here,$\delta(\cdot,\cdot)$is the Dirac function, corresponding to the covariance of the noise.

For the observation points $x_1, \dots, x_n$, their corresponding observation values $y_1, \dots, y_n$ follow a multivariate Gaussian distribution:

$$\mathbf{y} \sim \mathcal{N}(Z\beta, K + \sigma_n^2 I) \tag{9}$$

Here, $Z = [z_1, \dots, z_n]^T$ is the feature matrix, $K = [k(x_i, x_j)]_{n \times n}$is the kernel matrix, andI is the identity matrix. The joint probability density (likelihood function) is:

$$p(\mathbf{y}|Z, \beta, K, \sigma_n^2) = \frac{1}{(2\pi)^{n/2}|K + \sigma_n^2 I|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - Z\beta)^T (K + \sigma_n^2 I)^{-1} (\mathbf{y} - Z\beta)\right) \tag{10}$$

Take the log - likelihood (simplifying the constant term):

$$ln\ p(\boldsymbol{y}) = -\frac{1}{2} ln\ |K + \sigma_n^2 I| - \frac{1}{2} (\boldsymbol{y} - Z\beta)^T (K + \sigma_n^2 I)^{-1} (\boldsymbol{y} - Z\beta) \tag{11}$$

By maximizing the log - likelihood to learn the parameters $\{\beta, l, \sigma_n^2\}$, it is equivalent to minimizing the negative log - likelihood:

$$\mathcal{L} = \frac{1}{2} \ln\ |K + \sigma_n^2 I| + \frac{1}{2} (\mathbf{y} - Z\beta)^T (K + \sigma_n^2 I)^{-1} (\mathbf{y} - Z\beta) \tag{12}$$

Linear coefficient β: Take the derivative with respect to β, set the derivative to zero, and obtain the analytical solution:

$$\hat{\beta} = (Z^T (K + \sigma_n^2 I)^{-1} Z)^{-1} Z^T (K + \sigma_n^2 I)^{-1} \mathbf{y} \tag{13}$$

This formula is a weighted least squares solution, and the weights are jointly determined by the kernel matrix and the noise.

Kernel parameter l and noise σ_n^2: Since the kernel matrix K depends on l, numerical optimization (such as gradient descent, L - BFGS) is needed to minimize $\mathcal{L}$, Use the chain rule to calculate the derivatives of $\mathcal{L}$with respect to l and σ_n^2. Let the prediction point be x_*, and its corresponding feature be z_*, Construct the joint distribution:

$$\begin{bmatrix}\mathbf{y}\\ y_*\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}Z\beta\\ z_*\beta\end{bmatrix}, \begin{bmatrix}K+\sigma_n^2 I & \mathbf{k}_*\\ \mathbf{k}_*^T & k_{**}+\sigma_n^2\end{bmatrix}\right) \tag{14}$$

where $k_* = [k(x_1, x_*), \ldots, k(x_n, x_*)]^T$is the covariance vector, and $k_{**} = k(x_*, x_*)$is the auto - covariance. According to the properties of the multivariate Gaussian conditional distribution, the posterior distribution of $y_*$is:

$$y_*|\mathbf{y} \sim \mathcal{N}(\mu_*, \sigma_*^2) \tag{15}$$

where the posterior mean is:

$$\mu_* = z_*\beta + \mathbf{k}_*^T(K+\sigma_n^2 I)^{-1}(\mathbf{y} - Z\beta) \tag{16}$$

3. Data Analysis

The spectral dataset of samples is accessible for download at http://lib.stat.cmu.edu/ datasets/tecator. This dataset comprises n = 240 samples. Each sample encompasses the moisture, fat, and protein contents of meat, along with the absorption spectrum measured by a near - infrared spectrum analyzer within the wavelength range of 850–1050 nanometers (nm). With an interval of every two wavelength bands, a total of 100 absorption spectrum data points were recorded. To systematically assess the accuracy, robustness, and generalization capability of the model, this study employs the Monte Carlo simulation approach. The specific procedures are as follows: The dataset undergoes independent random partitioning 100 times. In each partitioning, samples are allocated to the training set and the test set at a 6:4 ratio (i.e., the sample size of the training set accounts for 60% of the total sample size, while the test set accounts for 40%). A model is constructed based on the training set from each partitioning, and the corresponding test set is utilized to verify the model's prediction performance. The prediction errors of each experiment (e.g., mean squared error, absolute error) are recorded.Over 100 repeated experiments, the mean of the errors serves as the core indicator for measuring the model's accuracy, reflecting the average prediction precision of the model across multiple experiments. The standard deviation of the errors acts as the key indicator for evaluating the model's robustness, characterizing the degree of performance fluctuation of the model under different sample partitioning scenarios. Specifically, a smaller standard deviation implies that the model is less affected by the random partitioning of data, indicating greater robustness.

The above - mentioned experimental design can effectively circumvent the interference of the randomness in single - sample partitioning on the model evaluation results. Through multiple repeated verifications at the statistical level, it comprehensively and objectively characterizes the comprehensive performance of the model, providing a reliable experimental basis for subsequent model comparison and optimization. Meanwhile, comparisons are made with methods such as Functional Principal Component Analysis + eXtreme Gradient Boosting (FPCA + XGBoost), Functional Principal Component Analysis + K - Nearest Neighbors (FPCA + KNN), Functional Principal Component Analysis + Elastic Net (FPCA + ElasticNet), Functional Principal Component Analysis + Gaussian Process Regression (FPCA + GPR), Functional Principal Component Analysis + Decision Tree (FPCA + TREE), Principal Component Analysis + eXtreme Gradient Boosting (PCA + XGBoost), Principal Component Analysis + K - Nearest Neighbors (PCA + KNN), Principal Component Analysis + Elastic Net (PCA + ElasticNet), Principal Component Analysis + Gaussian Process Regression (PCA + GPR), and Principal Component Analysis + Decision Tree (PCA + TREE), so as to verify the effectiveness and advantages of the proposed method.

First, based on the absorption spectrum data of the meat sample, as well as the contents of moisture and fat, we predict the protein content of the meat sample. Let y denote the protein content, Z_1 denote the moisture content, Z_2 denote the fat content, t denote the wavelength range within 850 - 1050 nanometers, and X(t) represent the relatively easily measurable spectral curve.

Table 1. Statistics of errors and standard deviations for protein prediction by various models

Model Name	Mean Squared Error	Standard Deviation - 1	Absolute Error	Standard Deviation - 2
FGPAR	0.2619	0.0850	0.3332	0.0539
FPCA+XGBoost	0.3495	0.1207	0.3903	0.0583
FPCA+KNN	0.3429	0.1186	0.3895	0.0620
FPCA+ElasticNet	0.7858	0.2033	0.7087	0.0932
FPCA+GPR	0.9121	0.1566	0.7847	0.0773
FPCA+TREE	0.4473	0.1523	0.4164	0.0665
PCA+XGBoost	1.0435	0.3448	0.7563	0.1306
PCA+KNN	1.0089	0.1778	0.7841	0.0796
PCA+ElasticNet	0.9041	0.1133	0.7712	0.0528
PCA+GPR	0.9794	0.1584	0.8514	0.0733
PCA+TREE	1.3958	0.4879	0.8786	0.1533

Table 1 presents the error and standard deviation performances of different models in the task of predicting protein content. Through comparison, it can be seen that the FGPAR model is significantly superior to other models in terms of mean squared error and absolute error metrics. Meanwhile, the standard deviation of FGPAR is at an extremely low level, reflecting that the dispersion degree of its prediction results is small. This indicates that in the scenario of protein content prediction, the FGPAR model has higher prediction accuracy and better result stability. Compared with models that integrate algorithms such as FPCA, PCA with XGBoost, KNN, etc., its advantages are quite prominent.

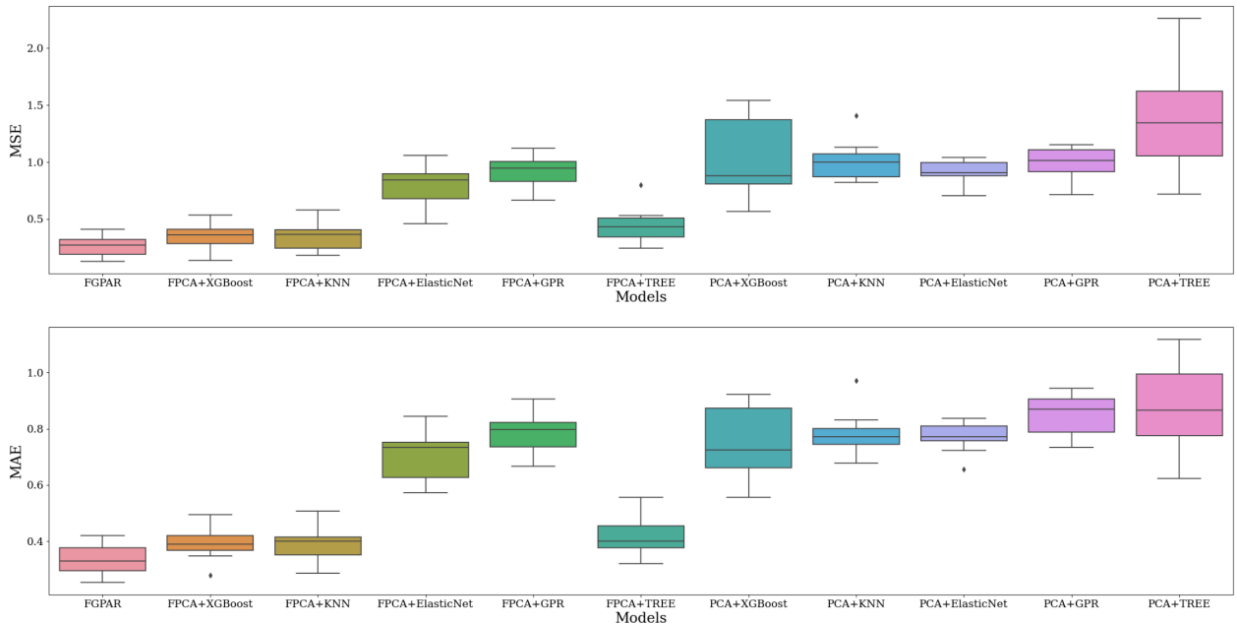

Figure 1. Boxplot of Protein Prediction Error

In the boxplot, the horizontal axis represents the models, while the vertical axis corresponds to the mean squared error (MSE) and mean absolute error (MAE), respectively. As shown in Figure 1 , the box corresponding to the FGPAR model exhibits the narrowest width and the lowest position, with

its median approaching the low - value interval. This indicates that its error distribution is highly concentrated, and the discreteness is extremely small. Conversely, for other models, the boxes are wider, the medians are higher, and the error fluctuations are significant. During multiple Monte Carlo simulations, the proportion of experiments in which the FGPAR model yields the minimum error is far higher than that of other models.

Again, we predict the water content of the meat sample using its absorption spectrum data, along with the moisture and fat contents. Let y denote the water content, Z_1 represent the protein content, Z_2 stand for the fat content, t define the wavelength range of 850–1050 nanometers, and X(t) signify the relatively easily measurable spectral curve.

Table 2. Statistics on Water Content Prediction Errors and Standard Deviations Across Different Models

Model Name	Mean Squared Error	Standard Deviation - 1	Absolute Error	Standard Deviation - 2
FGPAR	0.0261	0.0040	0.1122	0.0076
FPCA+XGBoost	0.1185	0.0504	0.2456	0.0358
FPCA+KNN	0.0896	0.0500	0.1972	0.0442
FPCA+ElasticNet	0.6588	0.3471	0.5891	0.1322
FPCA+GPR	0.9362	0.1800	0.7769	0.0809
FPCA+TREE	0.0749	0.0209	0.1893	0.0183
PCA+XGBoost	0.7118	0.1624	0.6113	0.0741
PCA+KNN	0.7072	0.1587	0.6473	0.0651
PCA+ElasticNet	0.6970	0.1893	0.6112	0.0704
PCA+GPR	0.9409	0.2396	0.8056	0.1145

Table 2 tabulates the performance of various models in the water - content prediction task. The data reveals that the mean squared error, mean absolute error, and their corresponding standard deviations of the FGPAR model all attain the minimum values across the entire dataset. In comparison with other models, the FGPAR model exhibits superior advantages in terms of prediction accuracy and result stability.

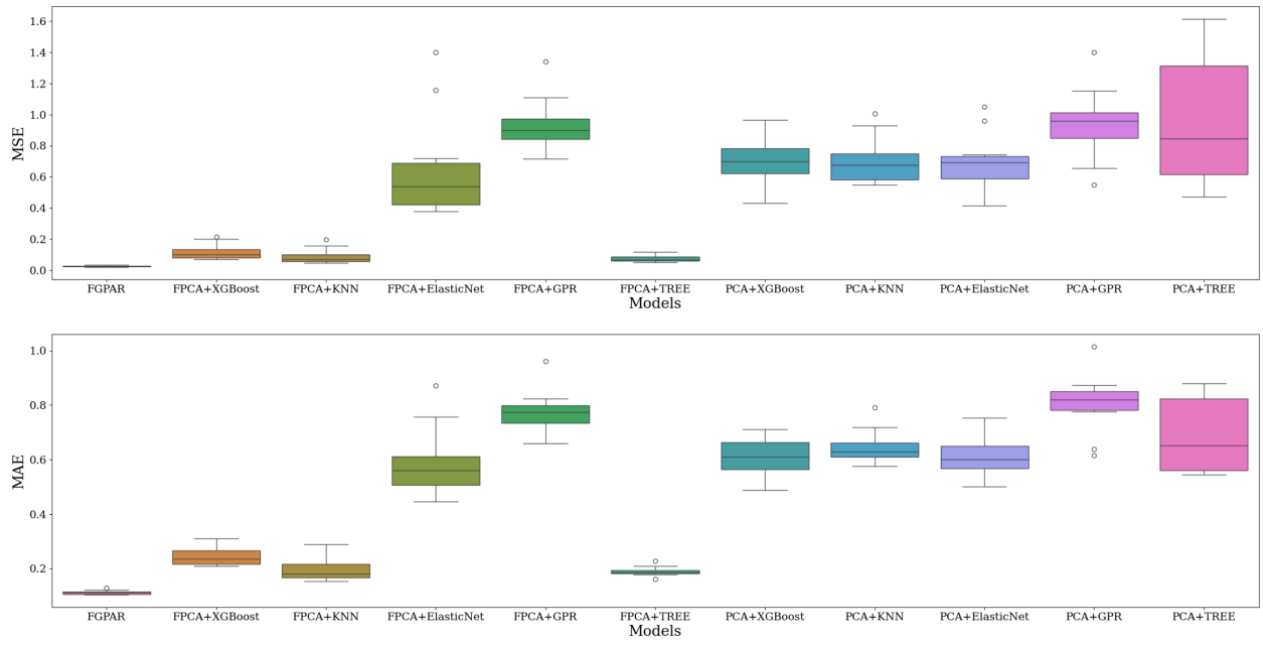

Figure 2. Boxplot of Water Prediction Error

In the box - plot, the box corresponding to the FGPAR model features a compact structure, free of outlier distribution, with errors concentrated in the extremely low - value interval. Conversely, as shown in Figure 2, the boxes of other models are wide, containing outliers and exhibiting a high degree of error dispersion. From the perspective of distribution characteristics, in water content prediction, the FGPAR model has a high proportion of small errors, and the predictability of its results is strong. In contrast, the errors of other models fluctuate significantly.

Finally, we predict the fat content of the meat sample based on the absorption spectrum data of the meat sample, along with the moisture and fat contents. Let y denote the fat content, Z_1 denote the protein content, Z_2 denote the water content, t denote the wavelength range of 850–1050 nanometers, and X(t) denote the spectral curve that is relatively easy to measure.

Table 3. Statistics of Errors and Standard Deviations for Fat Content Prediction by Different Models

Model Name	Mean Squared Error	Standard Deviation - 1	Absolute Error	Standard Deviation - 2
FGPAR	0.0269	0.0077	0.1136	0.0190
FPCA+XGBoost	0.0809	0.0399	0.2007	0.0371
FPCA+KNN	0.0549	0.0205	0.1724	0.0293
FPCA+ElasticNet	0.4908	0.2150	0.5231	0.0901
FPCA+GPR	0.8381	0.1336	0.7477	0.0607
FPCA+TREE	0.0762	0.0468	0.1796	0.0388
PCA+XGBoost	0.5886	0.2555	0.5323	0.1165
PCA+KNN	0.6813	0.2022	0.6085	0.0930
PCA+ElasticNet	0.6751	0.1369	0.6103	0.0613
PCA+GPR	0.7594	0.2549	0.6987	0.1484
PCA+TREE	0.9161	0.2707	0.6642	0.1224

Table 3 In the task of predicting fat content, the mean squared error, absolute error, and their respective standard deviations of the FGPAR model consistently achieve the best performance across all metrics. When compared with models integrated with algorithms like FPCA, PCA, XGBoost, and KNN, the FGPAR model maintains its advantages of high precision and strong stability. This verifies its general applicability in the task of predicting multiple components in meat.

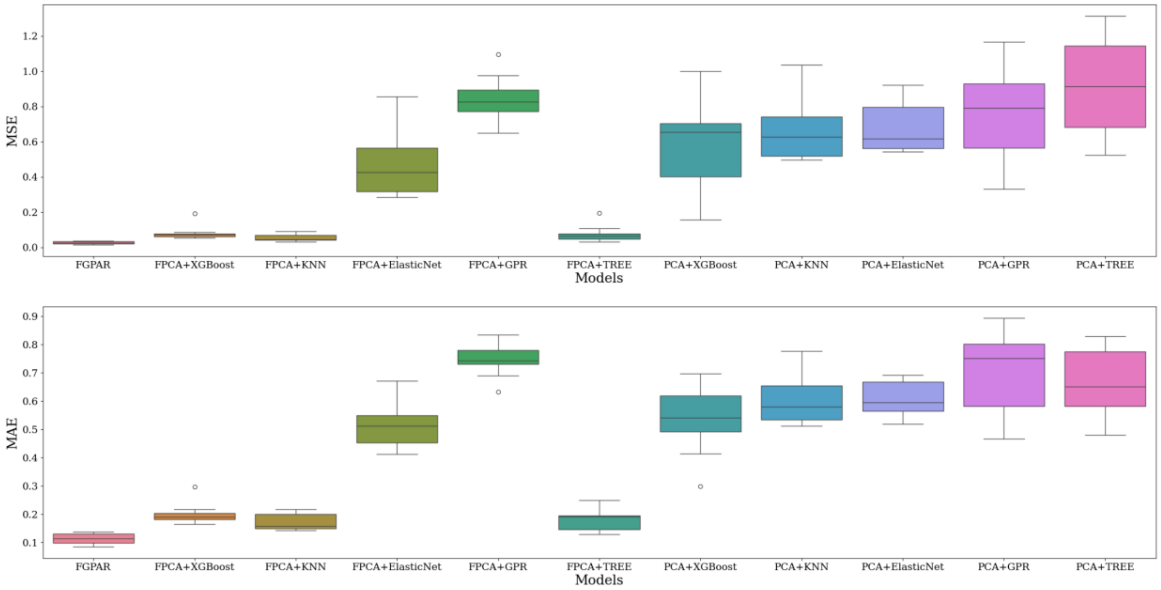

Figure 3. Boxplot of Fat Prediction Error

In the prediction of multiple components (protein, water, and fat), the error boxplot of the FGPAR model demonstrates a high degree of consistency: the box is narrow, the median is close to zero, and outliers are sparse. Taking fat prediction as an example, the error dispersion (box width) of other models is more than twice that of FGPAR. Moreover, the median deviation in water component prediction intuitively reflects its advantages. This stable performance across components indicates that FGPAR can more robustly capture the nonlinear associations between multiple indicators and covariates, avoiding error fluctuations caused by the interference of single - component features.Further inferences can be drawn from the pattern of the boxplot: As the proportion of functional data (e.g., time - series curves, spatial distribution functions) in the covariates increases and their forms become more complex, the advantages of FGPAR are likely to become more prominent. In comparative experiments, if the data contains functional features with high - frequency fluctuations, the error dispersion of other models will rise sharply. In contrast, due to the characteristics of FGPAR (i.e., the kernel function characterizes nonlinearity and the Gaussian process captures distribution), it can still maintain concentrated errors. This provides a visual verification basis for the application of this method in complex scenarios such as medical monitoring (e.g., modeling of continuous physiological indicators) and environmental simulation.

4. Conclusions

For the regression modeling task of mixed covariates that are functional - valued and vector - valued, this study proposes a partial functional nonparametric regression framework, along with a modeling strategy that integrates Gaussian processes and kernel methods. This approach characterizes the nonlinear dependence relationships of functional - type features through kernel functions, leverages Gaussian processes to capture the underlying probabilistic distribution properties of the data, and simultaneously integrates information from vector - valued features. In this way, a nonparametric regression model suitable for mixed data structures is constructed, enabling accurate prediction of the target variable (such as the content of meat components in this study).Empirical results show that, in the scenario of mixed data containing both functional - valued and vector - valued types, the prediction performance of the proposed method (measured by mean squared error, as well as the mean and standard deviation of absolute error) is significantly superior to that of traditional nonparametric methods and modeling approaches designed for a single data type. This verifies the adaptability and effectiveness of the proposed method for mixed data structures, providing a reliable solution for addressing the regression problems of mixed data that are functional - valued and vector - valued.

Although the method proposed in this study demonstrates advantages in regression tasks for mixed data, two aspects merit further in - depth exploration. First, the selection and parameter optimization of kernel functions exert a significant impact on model performance. In future research, the adaptive selection mechanism of kernel functions could be further investigated. By integrating properties such as the smoothness of functional - type features and the distribution characteristics of vector - valued features within the data, a dynamic kernel function combination strategy can be constructed. This aims to enhance the model’s adaptability to complex data structures. Second, the current model is based on the Gaussian process assumption. Going forward, it can be extended to a nonparametric regression framework for non - Gaussian processes (e.g., nonparametric models driven by Dirichlet processes or neural networks). Moreover, through multi - domain data experiments (such as mixed data of medical images and clinical indicators, as well as functional - type sensor data and vector - valued attribute data in environmental monitoring), the applicable boundaries of different models in mixed data scenarios can be compared. This will deepen the understanding of the generalization ability of nonparametric regression methods in mixed data modeling and provide methodological support for a broader range of practical problems.

References

[1] Wang Hongjian, Li Juan. Application and Progress of Bayesian Prior in Complex Data Modeling [J]. Statistics and Decision - Making, 2023, 39 (12): 34 - 38.

[2] Liu Hong, Chen Yang, Zhao Lei, et al. Construction of a Cardiovascular Disease Risk Assessment Model Based on Multi - Source Data Fusion [J]. Chinese Health Statistics, 2022, 39 (5): 678 - 681.

[3] Zhao Gang, Sun Ming. Research on the Application of Multi - Type Data Fusion in Regional Economic Growth Modeling [J]. Journal of Quantitative Economics and Technological Economics, 2023, 40 (7): 89 - 102.

[4] Huang Shanshan, Yang Jingping. Review of Functional Data Analysis Methods [J]. Statistics and Information Forum, 2010, 25 (12): 3 - 9.

[5] Yao Fang, Wu Xizhi. Estimation and Testing of Partial Functional Linear Models [J]. Journal of Mathematical Statistics and Management, 2018, 37 (2): 230 - 238.

[6] Zhang Ning, Wang Yan. Improvement of Partial Functional Linear Models and Their Application in Financial Data Modeling [J]. Research in Financial Economics, 2020, 35 (4): 115 - 126.

[7] Lin Hao, Zheng Min. Parameter Optimization and Application of Generalized Additive Models in the Collaborative Modeling of Mixed Data [J]. Journal of Quantitative Economics and Technological Economics, 2023, 40 (11): 156 - 171.

[8] Zhao meng, Wu Tao. Application of Bayesian Improved PFLM Model in Financial Mixed Data Prediction [J]. Statistics and Decision - Making, 2024, 40 (08): 89 - 93.

[9] Sun Yue, Zhou Ming. An Improved Method for Nonlinear Modeling of Functional Data Based on B - Spline Basis Functions [J]. Journal of Mathematical Statistics and Management, 2022, 41 (08): 1456 - 1468.

[10] Gao Xiang, Liu Jia. Nonlinear Correlation Modeling and Empirical Study of Mixed Data Driven by Multikernel Fusion [J]. Systems Engineering — Theory and Practice, 2023, 43 (13): 3421 - 3435.

[11] Chen Xi, Huang Wei. Modeling of Agricultural Mixed Data by Multikernel - Basis Function Fusion Under the Attention Mechanism [J]. Transactions of the Chinese Society of Agricultural Engineering, 2024, 40 (06): 201 - 209.