A Partial Functional Gaussian Process Regression Model Based on Additive Structure

Zunming Shen [#], Yijie Wang [#], Bowei Cui [#,*]

Nanjing University of Information Science and Technology, Nanjing, China, 210044

* Corresponding Author Email: Bowei_Cui@126.com

[#]These authors contributed equally.

Abstract. A partially functional linear model is a widely studied and used method whose response variables are related to both vector-valued and functional predictors. However, due to the constraints of linear structure, the model lacks some flexibility. Based on this, this paper proposes a partial functional Gaussian process regression model. On the one hand, the proposed method uses truncation rule to approximate the functional predictors. Then, the different predictors are mapped into a unified regenerated kernel Hilbert space (RKHS) for modeling using kernel techniques. By assigning a priori of additive structure to the Gaussian process and orthogonalizing the kernel function, this paper uses Sobol index to measure the importance of each prediction component's influence on the prediction of response variables. In addition to quantifying the uncertainty of predictions, the proposed method enables joint modeling of non-Euclidean predictors and response variables by selecting appropriate kernels for non-Euclidean spaces. Both simulation studies and empirical data analyses demonstrate that the proposed approach exhibits strong competitiveness compared to several classical predictive methods.

Keywords: Gaussian process, additive model, functional data analysis.

1. Introduction

With the rapid development of data collection and storage technologies, recording functional data has become increasingly convenient. However, traditional regression models encounter numerous challenges when handling such data [1]. On the one hand, these models fail to adequately address the issue of multicollinearity among predictors; on the other hand, they are often ineffective in capturing the intrinsic characteristics of functional data, leading to suboptimal or even failed modeling outcomes. Furthermore, in many practical applications, the factors influencing the response variable are not limited to functional data but may also involve other types of random variables. Therefore, developing joint regression models that can effectively handle mixed types of predictors and response variables has become a pressing challenge [2].

In this paper, we consider a generalized functional regression model, that is, the case where the predictor variables include functional random variables and vector-valued random variables, and the response variables are continuous scalars. Several studies have explored such regression models, including the partially functional linear model [3]. Zhang et al. [4] proposed an innovative approach to partially functional linear modeling. However, these models assume a linear relationship between predictors and the response variable, which limits their flexibility. Nonparametric statistical methods, which do not require a linear structure assumption, offer greater flexibility compared to linear models. Examples include ensemble learning models such as Random Forests [5], XGBoost [6], and Bagging [7]. However, these methods have certain limitations, such as not accounting for the intrinsic characteristics of functional data, as well as high computational complexity and storage costs. With the advancement of deep learning, regression methods based on neural networks have become increasingly mature. Models like Backpropagation (BP) neural networks and Long Short-Term Memory (LSTM) networks [8,9] can implicitly learn the intrinsic features of functional data and mitigate the curse of dimensionality through implicit regularization, making them effective and flexible semi-parametric models. However, in practical applications, regression models need not only flexibility but also interpretability, which neural networks generally lack. Additionally, uncertainty

estimation of the response variable is often crucial in predictive tasks, but the aforementioned models only provide point predictions. Gaussian process regression (GPR) models, through the flexible selection of kernel functions, can capture the nonlinear relationships between predictors and the response variable while providing uncertainty estimates for predictions. This capability has led to their widespread application in fields such as biomedicine [10], industrial manufacturing [11], and aerospace engineering [12]. By endowing Gaussian processes with an additive structure, these models achieve both flexibility and interpretability. In light of these considerations, this paper aim to extend the additive Gaussian process regression model to scenarios involving complex predictors—namely, functional and vector-valued predictors.

The aim of this paper is to propose a partially functional additive Gaussian process model for both functional and vector-valued predictors, focusing on enhancing the model's flexibility and interpretability. The key innovations of the proposed method are reflected in the following aspects: First, functional principal components are used to approximate the functional random variables, and kernel techniques are employed to map both functional and vector-valued predictors into a unified Reproducing Kernel Hilbert Space (RKHS) for joint modeling. Second, by adopting an additive Gaussian process regression framework, the method not only captures the complex relationships between predictors and the response variable effectively but also ensures model interpretability. Third, the Gaussian process regression provides uncertainty estimates for predictions, enhancing the reliability of the results. Lastly, with the appropriate selection of kernel functions, the model can handle not only vector-valued random variables but also non-Euclidean random variables. Additionally, when the kernel function is specified as a linear kernel, the proposed model reduces to a special case of the partially functional linear model. Extensive simulation experiments and empirical data analyses demonstrate that the proposed approach outperforms several classical regression models in terms of both prediction accuracy and robustness.

2. Theory and methods

Suppose that the observation $\{X_i(t), Z_i, y_i\}_{i=1}^n$ is generated by a partial functional regression model as follows

$$y_i = f\left(X_i(t), Z_i\right) + \xi_i \tag{1}$$

Where $X_i(t) \in \mathcal{L}^2$, $\mathcal{L}^2$ is a square integrable space, $Z_i = \left(z_{i1}, \cdots, z_{id}\right)^\top \in R^d$ R^d is a d-dimensional Euclidean space, $t \in \mathcal{T}$, $\mathcal{T}$ is a tight set, and ξ_i is a random variable that is independent of the predictor variables and obeys a normal distribution.

First, for the actual collected functional data, it is in the form of discrete points, so these discrete points need to be smoothed by the basis function. Specifically, this paper performs a Karhunen-Loève expansion of the stochastic function in terms of a functional principal component basis function as follows:

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \alpha_{ik} \varphi_k(t) \tag{2}$$

Where $\alpha_{ik} = \int_{t \in \mathcal{T}} \left(X_i(t) - \mu(t)\right) \varphi_k(t) dt$ is the principal component score, $\mu(t)$ is the mean function, and $\varphi_k(t)$ is the principal component basis function. Without loss of generality, assuming that the mean function is equal to 0, a truncated approximation of Eq. (2) can be obtained through the eigenvalue cumulative contribution criterion:

$$X_i(t) \approx \sum_{k=1}^{K} \alpha_{ik} \varphi_k(t) \tag{3}$$

Where K is the number of truncations? Next, since the functional principal component basis functions are a set of orthogonal basis functions, the L2 distances in the square productable space are isometric isomorphisms with the Euclidean distances, and all equations (1) can be approximated as equivalent to:

$$y_i = f\left(\alpha_{i1},\cdots,\alpha_{iK},Z_i\right)+\xi_i \tag{4}$$

Then, assuming that the connection function has a Gaussian process prior with additive structure, the proposed method can rewrite equation (4) as

$$\begin{aligned} f\left(\alpha_{i1},\cdots,\alpha_{iK},Z_i\right) &= f_1\left(\alpha_{i1}\right)+f_2\left(\alpha_{i2}\right)+\cdots+f_{K+1}\left(z_{i1}\right)+f_{K+2}\left(z_{i2}\right)+f_{K+d}\left(z_{id}\right) \\ &+f_{12}\left(\alpha_{i1},\alpha_{i2}\right)+\cdots+f_{12\ldots d+K}\left(\alpha_{i1},\alpha_{i1},z_{i1},\cdots,z_{id}\right) \end{aligned} \tag{5}$$

In the Gaussian process regression model with additive structure, the additive structure of the function decomposition is realized by the additive structure of the kernel, and the decomposition process of the additive kernel is shown as follows: first define a one-dimensional kernel function for each dimension $j=1,2,...,d+K$, define a one-dimensional kernel function, and from this class introduce 2nd order, 3rd order up to $d+K$ order kernel function.

$$\begin{aligned} k_1\left(s,s'\right) &= \sigma_1^2\sum_{j=1}^{K+d}k_j\left(s_j,s_j'\right) \\ k_2\left(s,s'\right) &= \sigma_2^2\sum_{j=1}^{K+d}k_j\left(s_j,s_j'\right)k_j\left(s_j,s_j'\right) \\ k_c\left(s,s'\right) &= \sigma_c^2\sum_{1\le j_1\le j_2\le\cdots\le j_d\le K+d}\left[\prod_{l=1}^{c}k_{j_l}\left(s_{j_l},s_{j_l}'\right)\right] \end{aligned} \tag{6}$$

Where $s_i=\left(\alpha_{i1},\cdots,\alpha_{iK},Z_i\right)$. The parameters σ_c control the relative importance of different output dimensions c in the function. Our model takes the same form as (5) only with the addition of a GP f_0 with a constant kernel, defining$[D]\coloneqq\{1,...,K+d\}$, and this paper stipulate that each f_j satisfies the orthogonality constraint:

$$\int_{\chi_j} f_j(s_j)p_j(s_j)ds_j = 0 \tag{7}$$

For $j\in[D]$, where χ_j and p_j are the sample space and density of the input feature s_j, respectively. This paper now construct kernels for each f_j. For each feature j that has the basic kernel k_j, it is known that under $S_j\coloneqq\int_{\chi_j} f_j(s_j)p_j(s_j)ds_j=0$, f is another GP with a modified kernel k_j.

$$f_j(\cdot)\left|\int f_j\left(s_j\right)p_j\left(s_j\right)ds_j=0\sim\mathcal{GP}\left(0,\tilde{k}_j\right)\right. \tag{8}$$

Where $\tilde{k}_j\left(s_j,s_j'\right)=k_j\left(s_j,s_j'\right)-\mathbb{E}\left[S_jf_j\left(s_j\right)\right]\mathbb{E}\left[S_j^2\right]^{-1}\mathbb{E}\left[S_jf_j\left(s_j'\right)\right]$, $\mathbb{E}\left[S_jf_j(\cdot)\right]=\int p_j\left(s_j\right)k_j\left(s_j,\cdot\right)ds_j$, $\mathbb{E}\left[S_j^2\right]=\iint p_j\left(s_j\right)p_j\left(s_j'\right)k_j\left(s_j,s_j'\right)ds_jds_j'$

This paper calls k_j the kernel of constraints and for higher-order interaction terms, the constraints satisfy $\int_{\chi_j} f_j(s_j)p_j(s_j)ds_j=0\forall j\in u$ where $s_u\coloneqq\{s_j\}_{j\in u}$. Then the Functional ANOVA method is used to decompose the function $f(s)$ into the form $f(s)=\sum_{u\subseteq[D]}f_u(s_u)$. Applying the FANOVA decomposition to the OAK construction, it can be found that the functions considered in OAK are exactly the components of the FANOVA decomposition. The orthogonality of OAK leads to the homogeneity of ANOVA.

$$R\coloneqq\mathbb{V}_x[f(s)]=\sum_{u\subseteq[D]}R_u, \tag{9}$$

Where $R_u\coloneqq\mathbb{V}_x[f_u(s)]$ is defined as the Sobol exponent of the feature set u. Finally, we obtain the Sobol index of the associated posterior mean function:

$$\tilde{R}_u = \frac{\mathbb{V}_s\left[m_u(s)\right]}{\mathbb{V}_s[m(s)]} \tag{10}$$

Where m_u and m are defined as the posterior mean functions of f_u and f respectively.

$$m_u(s) = \sigma^2_{|u|}\left(\sum_{j\in u}\tilde{k}_j\left(s_j, S_j\right)\right)K(S,S)^{-1}\mathbf{y} \tag{11}$$

$$\begin{aligned}\mathbb{V}_s\left[m_u(s)\right] &= \sigma^4_{|u|}y^{\top}K(S,S)^{-1}\sum_{j\in u}\\ &\left(\int \tilde{k}_j\left(s_j, S_j\right)\tilde{k}_j\left(s_j, S_j\right)^{\top}\mathrm{d}p_j\left(s_j\right)\right)K(S,S)^{-1}\mathbf{y}\end{aligned} \tag{12}$$

Where $K(S,S)^{-1}$ denotes the training input covariance for all inputs, S_j and y denote the j th column of S and the vector of output observations, and $\sigma^2_{|u|}$ is the correlation covariance order of the order u interaction.

3. Data analysis

3.1. Simulation experiments

In order to comprehensively assess the prediction performance of the proposed method, we consider simulations under different experimental settings. Specifically, we consider the cases with different sample sizes, connection functions, and variable distributions. For comparison methods, we choose to use LASSO regression, Ridge regression, and Random Forest and Xgboost models. For the evaluation indicator, we choose to use the Root Mean Squared Error (RMSE). For the training and testing process, we choose to use 70% of the samples as the training set to train the data, and 30% of the data as the test set to test the models. In order to truly reflect the differences between models, this paper uses Monte Carlo simulation of multiple repeated tests, and then the mean and standard deviation of the evaluation indicator of multiple tests are used as the basis of discrimination. Next, the model comparisons are first considered for different numbers of samples, assuming that n is the number of samples, and the connection function between the predictor and response variables is constructed as follows:

$$Y_0 = 0.1\cdot\sin(2\pi X_1) + 0.1\cdot X_2^2 + 0.2\cdot X_3\cdot X_4 + \sin\left(\frac{\pi}{2}Z_1\right) + \cos\left(\frac{3\pi}{4}(Z_2+Z_3)\right) + 0.01\cdot\mathrm{rnorm}(n) \tag{13}$$

Where X is the principal component score of the function-type data, Z is the vector-valued random variable, the error term obeys the standard normal distribution, and the pre-smoothing about the function-type data is carried out using the Fourier basis function. The specific results are shown in Table 1, 2 and Figure 1, 2 below:

Table 1. Comparison results with different sample size conditions.
(All considering functional information).

MODEL	$n=100$	$n=150$	$n=200$	$n=300$
FOAK	0.3054 (0.0464)	0.2186 (0.0363)	0.1405 (0.0243)	0.0894 (0.0154)
FPCA+LASSO	0.7829 (0.0895)	0.8339 (0.0445)	0.8559 (0.0564)	0.7438 (0.0404)
FPCA+Ridge	0.8482 (0.0905)	0.8967 (0.0369)	0.9294 (0.0582)	0.8508 (0.0341)
FPCA+RF	0.7788 (0.0587)	0.6475 (0.0535)	0.6875 (0.0459)	0.6505 (0.0485)
FPCA+Xgboost	0.8257 (0.0934)	0.6850 (0.0654)	0.6813 (0.0472)	0.6897 (0.0548)

Note: Means of evaluation indicator outside parentheses and standard deviations of evaluation indicator inside parentheses.

Table 2. Comparison results with different sample size conditions (not considering functional information in the comparison methods).

MODEL	$n=100$	$n=150$	$n=200$	$n=300$
FOAK	0.3021 (0.0381)	0.2099 (0.0324)	0.1301 (0.0147)	0.1088 (0.0295)
LASSO	0.7788 (0.1257)	0.8303 (0.0951)	0.8726 (0.0592)	0.7895 (0.0276)
Ridge	1.5399 (0.0998)	1.5454 (0.1841)	1.4280 (0.1169)	1.0465 (0.0468)
RF	0.7791 (0.1198)	0.6770 (0.0513)	0.7207 (0.0469)	0.6657 (0.0228)
Xgboost	0.8640 (0.1219)	0.6892 (0.0692)	0.7568 (0.0438)	0.7090 (0.0222)

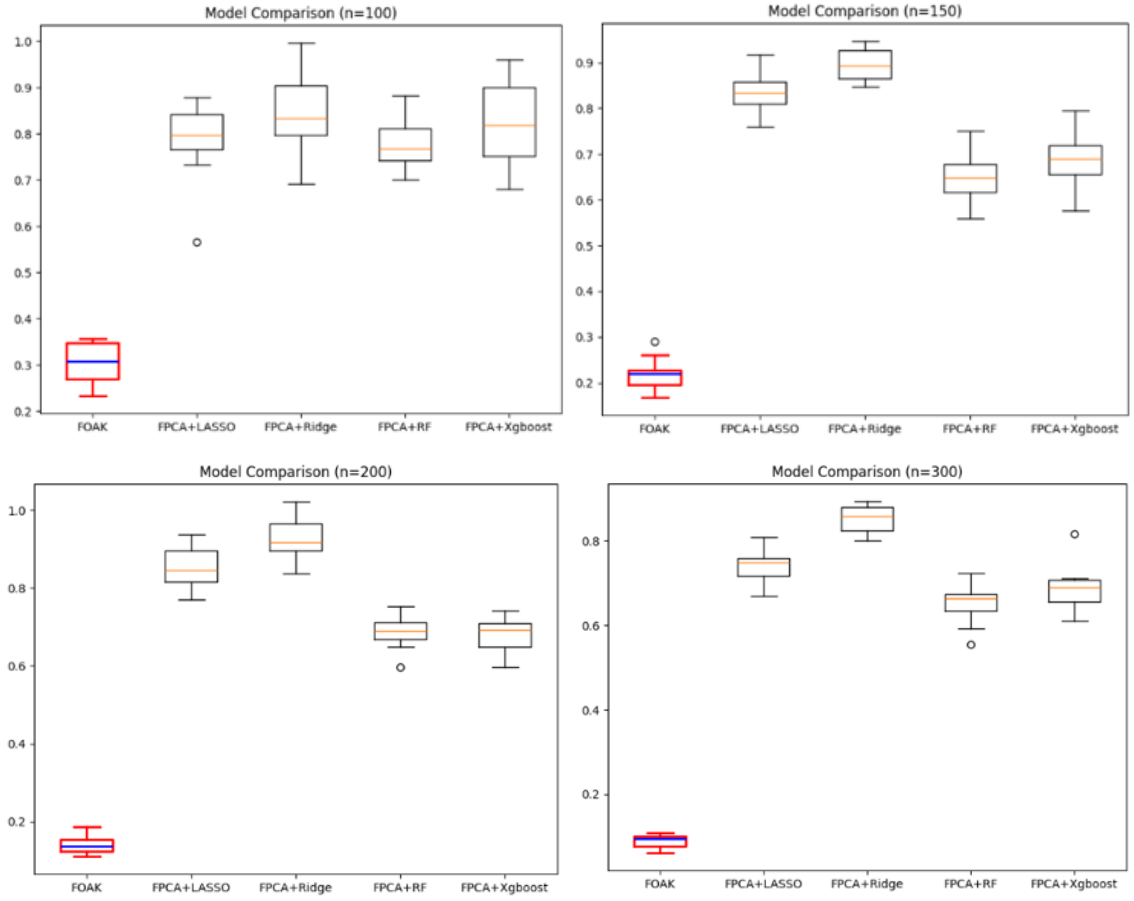

Figure 1. Distribution of comparison results with different number of sample (all considering functional information).

In this section, we analyze FOAK against the comparison methods under different sample number conditions, which demonstrates its excellent accuracy and stability. Specifically, at a sample size of 100, the average RMSE of the FOAK method is reduced by about 61% compared to that of the LASSO method that takes functional information into account, and 64%, 61% and 63% compared to that of the Ridge, RF and Xgboost methods that take functional information into account, respectively. In addition, we find that as the sample size increases, the error of the FOAK model becomes smaller and more convergent, and the advantage over the comparison methods becomes more obvious. When the sample size reaches 300, the average RMSE of the FOAK method has been reduced by more than 86% compared with the other methods which shows that the model has a very high accuracy. At the same time, the standard deviation of RMSE of the FOAK model is also the smallest among multiple methods, showing its excellent stability. In addition, the advantage of our method over other comparative methods when functional information is not considered may be more obvious. The advantages of FOAK over other methods can be more clearly observed through box plots. And the superiority of our model may mainly due to its utilization of function information and the flexibility of the model fitting function is again enhanced by additive Gaussian process regression.

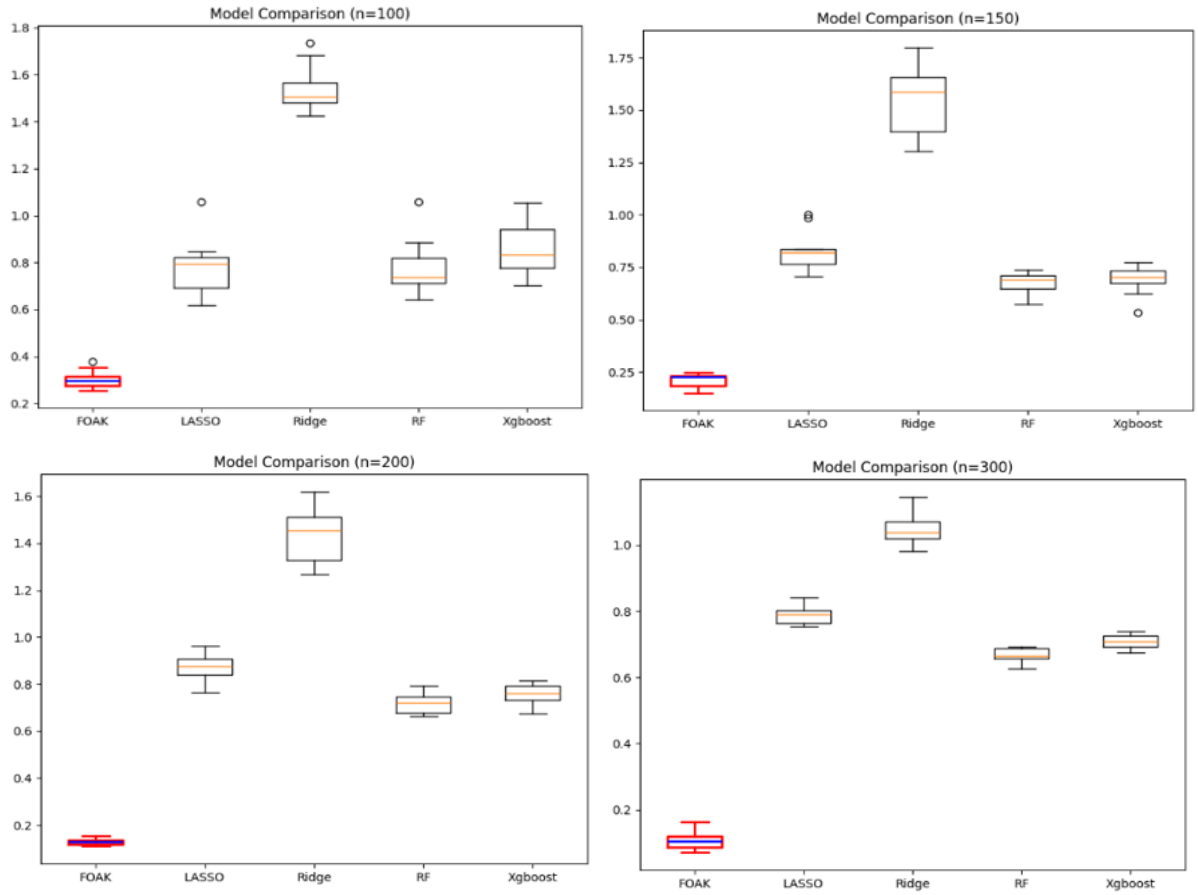

Figure 2. Distribution of comparison results with different number of samples.
(Not considering functional information in the comparison methods).

Next, this paper considers the prediction performance of the proposed method with different sample distributions. Where the number of samples is set as$n = 100$, the connection function between the predictor and response variables is the same as that in case 1.

$$Y_0 = 0.1 \cdot \sin(2\pi X_1) + 0.1 \cdot X_2^2 + 0.2 \cdot X_3 \cdot X_4 + \sin\left(\frac{\pi}{2} Z_1\right) + \cos\left(\frac{3\pi}{4}(Z_2 + Z_3)\right) + 0.01 \cdot \mathrm{rnorm}(n) \tag{14}$$

The results are shown in Table 3, 4 and Figure 3, 4:

Table 3. Comparison results with different sample distributions.
(All considering functional information)

MODEL	$a_k \sim 1*N(0,1)$	$a_k \sim 0.5*N(0,1)$	$a_k \sim 0.1*N(0,1)$	$a_k \sim 0.01*N(0,1)$
FOAK	0.3054 (0.0464)	0.1975 (0.0531)	0.1759 (0.0353)	0.2332 (0.0503)
FPCA+LASSO	0.7829 (0.0895)	0.7922 (0.0825)	0.8085 (0.0959)	0.7966 (0.0979)
FPCA+Ridge	0.8482 (0.0905)	0.8708 (0.0766)	0.8338 (0.0863)	0.8310 (0.0845)
FPCA+RF	0.7788 (0.0587)	0.7381 (0.0716)	0.6846 (0.0630)	0.6596 (0.0460)
FPCA+Xgboost	0.8257 (0.0934)	0.7824 (0.0671)	0.7221 (0.0899)	0.7118 (0.0812)

Table 4. Comparison results with different sample distributions.
(Not considering functional information in the comparison methods)

MODEL	$a_k \sim 1*N(0,1)$	$a_k \sim 0.5*N(0,1)$	$a_k \sim 0.1*N(0,1)$	$a_k \sim 0.01*N(0,1)$
FOAK	0.3021 (0.0381)	0.1958 (0.0420)	0.1819 (0.0341)	0.1862 (0.0580)
LASSO	0.7788 (0.1257)	0.8264 (0.0820)	0.7640 (0.0903)	0.8602 (0.0536)
Ridge	1.5399 (0.0998)	1.1624 (0.0799)	0.8620 (0.0828)	0.9027 (0.0687)
RF	0.7791 (0.1198)	0.6837 (0.0612)	0.7285 (0.0705)	0.7534 (0.0664)
Xgboost	0.8640 (0.1219)	0.7161 (0.0926)	0.7553 (0.0783)	0.8096 (0.0746)

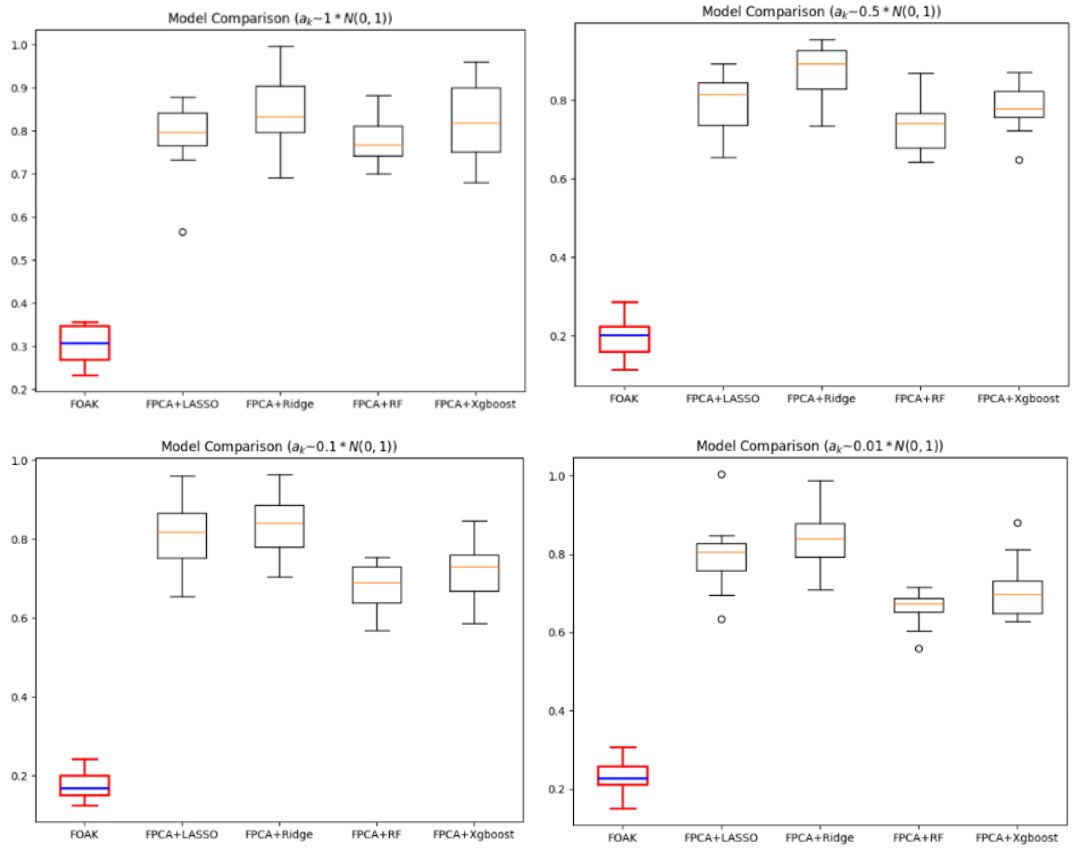

Figure 3. Distribution of comparison results with different sample distributions (all considering functional information).

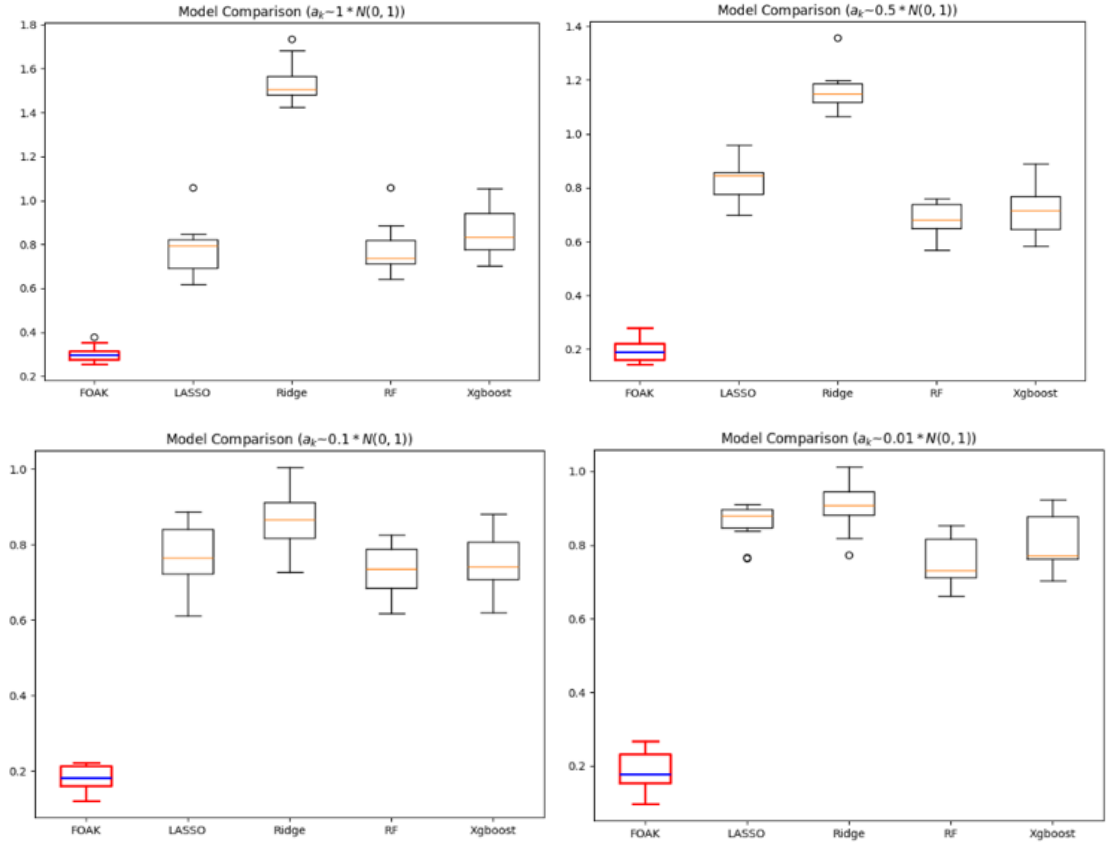

Figure 4. Distribution of comparison results with different sample distributions (not considering functional information in the comparison methods).

In this section, we analyze in detail the prediction performance of the FOAK method with different sample distributions. Different sample distributions determine the fluctuation levels of the data, so understanding and coping with these fluctuations is important for assessing the performance of the model. Regardless of the data fluctuations, the FOAK method demonstrates extremely high accuracy and stability. Specifically, when $a_k \sim 1*N(0,1)$, the average RMSE of FOAK method is significantly lower than that of the LASSO, Ridge, RF, and Xgboost methods, which take function information into account. When the level of data fluctuation decreases, the FOAK method still maintains high accuracy, and the advantage over the other methods is even more obvious, with the average RMSE decreasing by more than 78% at one point, demonstrating an extremely strong ability to capture information. In addition, the standard deviation of the RMSE of the FOAK method is also smaller under a variety of sample distributions, showing its excellent stability with various fluctuations in the data.

Finally, a different connection function case is considered, setting the number of samples $n = 100$ and the sample distribution $a_k \sim 1*N(0,1)$. In addition to the connection function already used above, another connection function is considered in this paper as shown below:

$$Y_1 = 0.1 \cdot X_1 \cdot X_2 + 0.2 \cdot log(X_3 + 1) + 0.2 \cdot X_4^3 + sin\left(\frac{\pi}{3}(Z_1 + Z_2 + Z_3)\right) + 0.01 \cdot rnorm(n) \quad (15)$$

The results are shown in Table 5-6 and Figure 5-6:

Table 5. Comparison results with different connection functions (all considering functional information).

MODEL	$y = Y_0$	$y = Y_1$
FOAK	0.3054 (0.0464)	0.0989 (0.0366)
FPCA+LASSO	0.7829 (0.0895)	0.6519 (0.0499)
FPCA+Ridge	0.8482 (0.0905)	0.7255 (0.0518)
FPCA+RF	0.7788 (0.0587)	0.4578 (0.0386)
FPCA+Xgboost	0.8257 (0.0934)	0.4663 (0.0769)

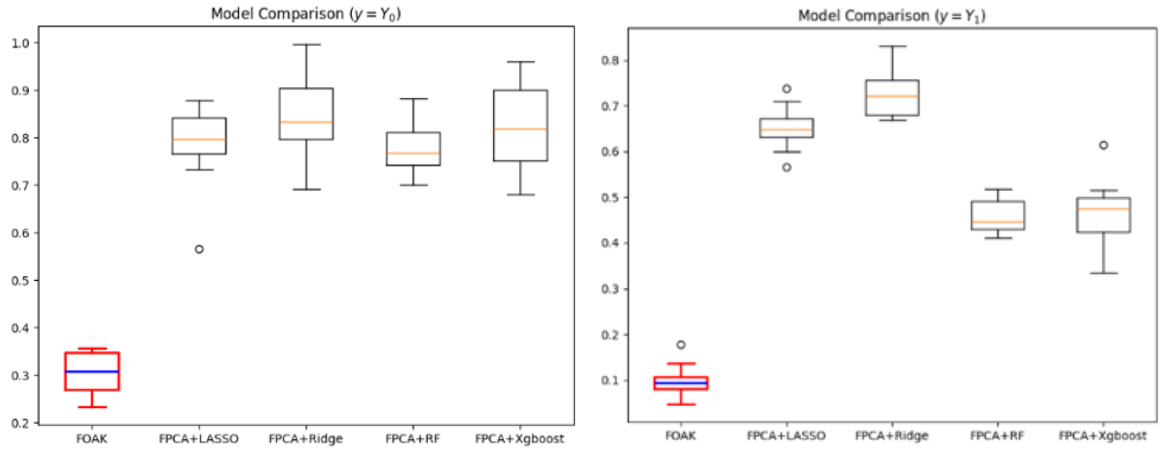

Figure 5. Distribution of comparison results with different connection functions. (All considering functional information).

Table 6. Comparison results with different connection functions. (Not considering functional information in the comparison methods).

MODEL	$y = Y_0$	$y = Y_1$
FOAK	0.3021 (0.0381)	0.1090 (0.0428)
LASSO	0.7788 (0.1257)	0.6453 (0.0607)
Ridge	1.5399 (0.0998)	0.8328 (0.0625)
RF	0.7791 (0.1198)	0.5254 (0.0529)
Xgboost	0.8640 (0.1219)	0.5393 (0.0779)

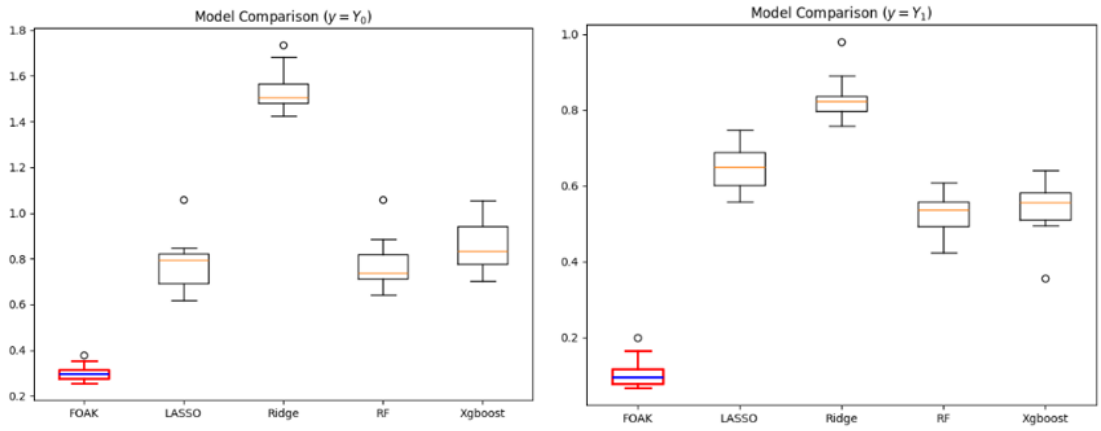

Figure 6. Distribution of comparison results with different connection functions. (Not considering functional information in the comparison methods).

Different connection functions may have impact on the results of the model, so it is necessary to analyze the performance of the model with respect to different connection functions. In this section, we analyze the performance of the FOAK method in detail with different connection functions and the FOAK method still demonstrates its high accuracy and stability. Specifically, for different functions, the average RMSE of the FOAK method are all significantly lower than those of the LASSO, Ridge, RF, and Xgboost methods that take functional information into account. The advantages of FOAK are obvious for the connection function (Y_0), which performs poorly for the comparison methods, and even more significant for the connection function (Y_1), which works well for the comparison methods, where the average RMSE is reduced by more than 86% compared to the Ridge model which takes into account the functional information, and about 80% compared to the other methods, indicating an extremely strong flexibility, perhaps also thanks to the application of additive Gaussian process regression and considering nonlinearity of f. In addition, the standard deviation of the RMSE of the FOAK method is minimal under a variety of connection functions, which shows the excellent stability of the method in different connection conditions. The box plots provide a clearer view of the superior performance of the FOAK method, maintaining high accuracy and stability regardless of changes in the connection functions.

3.2. Real data analysis

The first dataset is the meat dataset (https://lib.stat.cmu.edu/datasets/). This dataset has 240 samples, each containing the moisture, fat and protein of the meat, as well as the absorption spectra measured by a near-infrared (NIR) spectral analyzer in the wavelength range of 850-1050 nanometers (nm). A total of 100 absorption spectra data were recorded at intervals of two bands. This dataset has also been used in the literature for other functional data. Our aim is to predict the protein of a meat sample based on the absorption spectrum data of the meat sample, as well as the moisture and fat. Denote Y as protein content, Z_1 as water content, Z_2 as fat content, and t as values in the 850-1050 nm band, $X(t)$ which is a more easily measured spectral profile to analyze the chemical composition of the meat. In order to compare the performance of FOAK as well as the four comparison methods in whether or not considering the functional information as mentioned earlier, we randomly divided the data with a total sample size of 240 into two parts of the training set and the test set by 70% and 30%, and the results are shown in Tables 7.

From the mean and standard deviation of prediction RMSE in Table 7 and Figure 7-8, the optimal method is FOAK, which has smaller mean and standard deviation of prediction RMSE than the other four methods. The advantage of FOAK over the comparison methods that do not take into account the information of the function is more obvious, thus proving its excellent performance in practical application. The superior performance of the FOAK method with meat data can also be observed more visually through box plots.

Table 7. Comparison results for the meat dataset.

MODEL	μ (functional)	σ (functional)	μ (non-functional)	σ (non-functional)
FOAK	0.6493	0.0707	0.6464	0.0712
FPCA+LASSO	1.0283	0.0878	1.0457	0.1311
FPCA+Ridge	3.0908	0.1583	3.2523	0.2872
FPCA+RF	0.7537	0.0834	0.8277	0.1239
FPCA+Xgboost	0.7421	0.1173	0.8436	0.1246

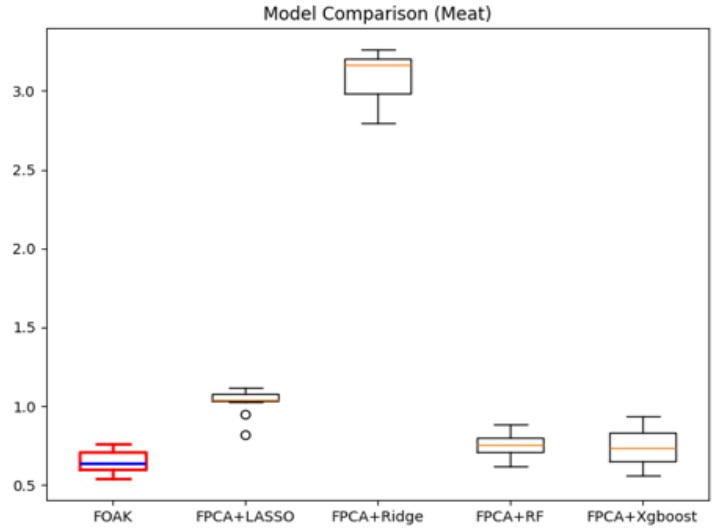

Figure 7. Distribution of comparison results for meat dataset. (All considering functional information).

The feature importance results for the meat dataset in Table 8 show that the 5th feature (Fat) has the highest importance value (0.6792), indicating that it has a key role in predicting the results. This is followed by the 6th feature (Water) which has an importance value of 0.0890 and also has an impact on the model. The first, third, and fourth principal component features have importance values of 0.0485, 0.0371, and 0.0275, respectively, which are relatively small but still influential. In addition, interaction effects such as Fat and Water also show the importance of interaction in the model, especially [Fat, Water] which has an importance value of 0.0262. Overall, Fat has a much greater impact on the dataset than any other feature and contributes the most to the prediction results. By analyzing the importance of these features, we are able to better understand the key factors that influence the prediction.

Table 8. Results of feature importance for meat dataset.

INDEX	VALUE	INDEX	VALUE
[4]	0.6792	[1, 5]	0.0204
[5]	0.0890	[1, 4]	0.0124
[0]	0.0485	[1, 3]	0.0116
[2]	0.0371	[1, 2]	0.0107
[3]	0.0275	[1]	0.0094
[4, 5]	0.0262	[0, 1]	0.0073

Based on the results in Tables 9 and Figure 9-10, we actually find that for the other four comparison methods, whether or not the functional information is considered does not have a significant impact on the results, perhaps because this part of the data does not reflect strong functional properties. When compared with the other methods, the mean and standard deviation of the RMSE for the FOAK method are significantly lower than the other methods, indicating that the FOAK method has a clear advantage in accuracy and stability when fitting as well as predicting the data.

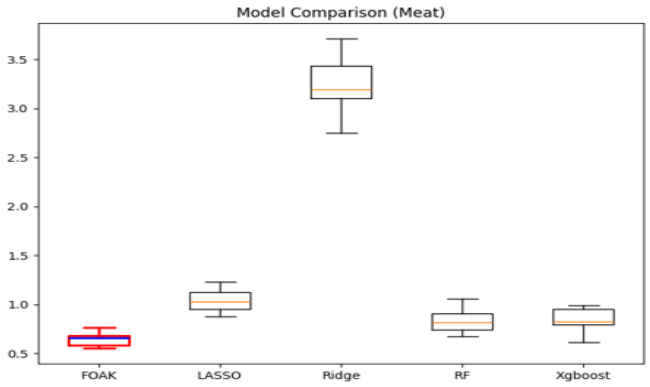

Figure 8. Distribution of comparison results for meat dataset.
(Not considering functional information in the comparison methods).

Table 9. Comparison results for corn dataset.

MODEL	μ (functional)	σ (functional)	μ (non-functional)	σ (non-functional)
FOAK	0.2746	0.0375	0.2957	0.0315
FPCA+LASSO	0.6691	0.0596	0.6684	0.0644
FPCA+Ridge	0.7942	0.1017	0.8025	0.0947
FPCA+RF	0.3370	0.0525	0.3342	0.0361
FPCA+Xgboost	0.3737	0.0419	0.3517	0.0322

Table 10. Results of feature importance for corn dataset.

INDEX	VALUE	INDEX	VALUE
[6]	0.7056	[2, 6]	0.0143
[0]	0.0919	[2, 4]	0.0084
[2]	0.0657	[0, 3]	0.0072
[5]	0.0385	[4, 6]	0.0060
[3, 6]	0.0231	[3, 4]	0.0057
[4]	0.0221	[3]	0.0053

The results of feature importance for the meat dataset in Table 10 show that the 7th feature (Protein) has the highest importance value (0.7056), indicating its key role. This is followed by the 1st feature (first functional principal component) with an importance index of 0.0919, which also has an impact on the results of the model. In addition, the interaction effects such as the third functional principal component with protein and moisture with protein also show the importance of the interaction in the model. Overall, the 7th feature had a much greater impact on the dataset than the other features and contributed the most to the results.

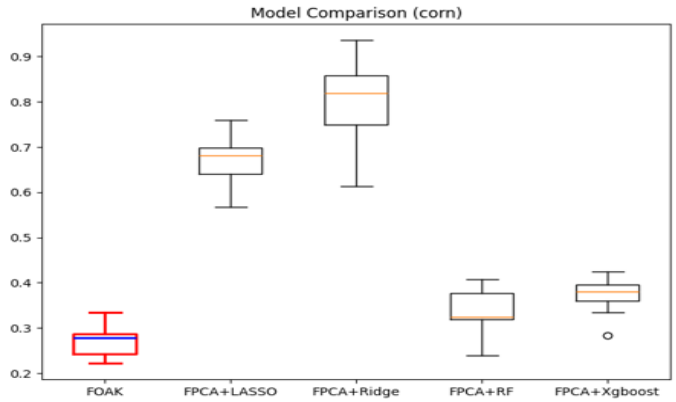

Figure 9. Distribution of comparison results for the corn dataset.
(All considering functional information).

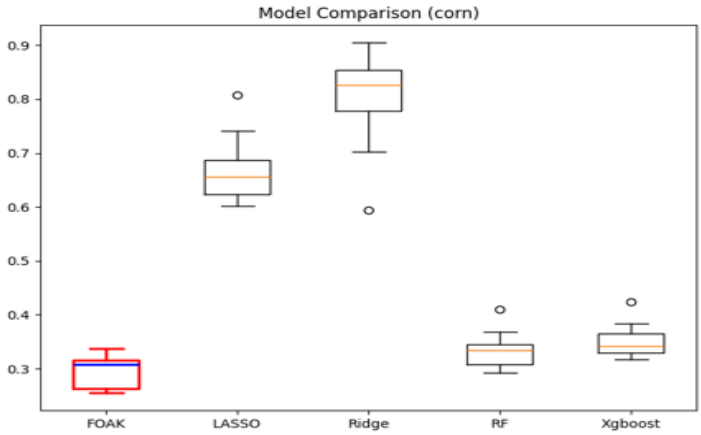

Figure 10. Distribution of comparison results for the corn dataset.
(Not considering functional information in the comparison methods)

4. Conclusion

In this paper, the proposed method develops a partially functional Gaussian process regression model based on an additive structure. This model effectively handles various types of covariates within a unified framework, including but not limited to vector-valued and functional data. Specifically, functional principal component analysis (FPCA) is first employed to approximate random functions, and kernel techniques are utilized to jointly model all predictors within a Reproducing Kernel Hilbert Space (RKHS), thereby ensuring the flexibility of the predictive model. Furthermore, by assigning an additive Gaussian process prior to the predictive function, we are able to decompose the contributions of different covariates, providing deeper insights into the relationships between the predictors and the response variable. Extensive simulation studies and real-world data analyses demonstrate that the proposed model outperforms traditional methods in comparison, exhibiting superior performance.

For future research directions, first, if covariates are observed in non-Euclidean spaces such as Riemannian manifolds, it will be necessary to introduce specific kernel functions suitable for these spaces to respect the intrinsic geometric structure of the data. Second, we can extend the proposed method to handle discrete response variables to address classification tasks. Lastly, our model can be further expanded to jointly predict multiple response variables, transforming it into a multi-output model capable of simultaneous predictions, which will broaden its applications in the field of multi-task learning.

Reference

[1] Hussain, J. (2020). High dimensional data challenges in estimating multiple linear regression. Journal of Physics: Conference Series, 1591.

[2] Tran, A., Maupin, K., Mishra, A., Hu, Z., & Hu, C. (2023). Additive Gaussian Process Regression for Metamodeling in High-Dimensional Problems. Volume 2: 43rd Computers and Information in Engineering Conference (CIE).

[3] Yu, P., Du, J., & Zhang, Z. (2020). Single-index partially functional linear regression model. Statistical Papers, 61, 1107-1123.

[4] Zhu, R., Zhou, G., &Zhang, X. (2018). Optimal Model Average Estimation for Partial Functional Linear Models. System Science and Mathematics Sciences, 38(07):777-800.

[5] Schonlau, M., & Zou, R. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20, 29 - 3.

[6] Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. International Journal of Distributed Sensor Networks, 18.

[7] Pham, H., & Ólafsson, S. (2019). Bagged ensembles with tunable parameters. Computational Intelligence, 35, 184 - 203. Li, X., Yuan, C., & Shan, B. (2020). System Identification of Neural Signal Transmission Based on Backpropagation Neural Network. Mathematical Problems in Engineering, 2020, 1-8.

[8] Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Computation, 31, 1235-1270.

[9] Wu, R., & Wang, B. (2018). Gaussian process regression method for forecasting of mortality rates. Neurocomputing, 316, 232-239.

[10] Kopsiaftis, G., Protopapadakis, E., Voulodimos, A., Doulamis, N., & Mantoglou, A. (2019). Gaussian Process Regression Tuned by Bayesian Optimization for Seawater Intrusion Prediction. Computational Intelligence and Neuroscience, 2019.

[11] Buelta, A., Olivares, A., Staffetti, E., Aftab, W., & Mihaylova, L. (2021). A Gaussian Process Iterative Learning Control for Aircraft Trajectory Tracking. IEEE Transactions on Aerospace and Electronic Systems, 57, 3962-3973.

[12] Pfingstl, S., Braun, C., Nasrollahi, A., Chang, F., & Zimmermann, M. (2022). Warped Gaussian processes for predicting the degradation of aerospace structures. Structural Health Monitoring, 22, 2531 - 2546.