Partial Linear Additive Integrated Regression Models Based on Functional Gaussian Processes

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

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

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

* Corresponding Author Email: 18771171720@163.com

Abstract. In this paper, proposing a partially linear additive Gaussian process-based integration model for regression problems containing functional-type and vector-valued predictor variables. The model downscales high-dimensional functional data by functional principal component analysis, extracts low-dimensional principal component scores, and constructs additive Gaussian process sub-models, i.e., the non-linear relationship between functional predictor variables and response variables is fitted by independent Gaussian process components, and the vector-valued predictor variables are selected by the linear portion of Lasso regularisation for variable selection. To enhance the model generalisation ability, the predictions of each sub-model are adaptively fused using the Stacking integration strategy. Simulation experiments show that the proposed method can effectively identify irrelevant vector-type variables and significantly outperforms the traditional functional linear model in terms of prediction error; real data analysis further validates its competitiveness in prediction tasks. In addition, the framework proposed in this paper is highly scalable: the feature representation of functional variables can be optimized by introducing multi-kernel learning, and the computational complexity can be reduced by combining with sparse Gaussian process, which can be extended to dynamic data flow modelling or combined with deep integration methods in the future, to provide a flexible and interpretable modelling paradigm for complex and heterogeneous data.

Keywords: gaussian process regression; partial linear models; functional principal component analysis; Stacking integration.

1. Introduction

With the rapid development of modern data acquisition technology, functional data has gradually become an important research object in the fields of biostatistics, medical diagnosis and chemometrics. Typical examples of such data, which are presented in the form of continuous functions, include spectral curves of substances, dynamic electrocardiogram signals, and real-time monitoring data of industrial processes. Multiple linear regression and logistic regression models in the framework of traditional regression analysis face significant theoretical limitations when dealing with functional data due to their inherent linear assumptions and strict requirements on sample independence. This limitation mainly stems from two aspects: first, the linear model is difficult to capture the nonlinear mapping relationship that may exist between the functional data and the response variable; second, the assumption of independent and homogeneous distribution is in essential conflict with the inherent temporal correlation characteristics of the functional data. This disconnects between theory and practice has led to the inability of traditional modeling methods to effectively explore the deep information embedded in functional data, which in turn affects the accuracy and predictive efficacy of statistical inference.

Smoothing is a key step to eliminate noise interference in functional data preprocessing.Wan Anis Farhah Wan Amir et al. proposed a flexible smoothing method based on β-splines, which enhances the model adaptability by introducing two shape parameters[1] and optimizes the smoothing parameters by combining roughness penalty and generalized cross validation (GCV), and is able to efficiently deal with the non-stationary time series and local fluctuation curves; Yang Ying, Yao Fang et al. further proposed an online estimation algorithm for real-time data streaming scenarios, which significantly improves the processing efficiency of high-dimensional functional-type data by

updating the mean and covariance functions through dynamic window widths[2]. Subsequently, Functional Principal Component Analysis (FPCA) has been widely used as a core tool for dimensionality reduction and feature extraction. sunnyG.W.Wang et al. developed an adaptive kernel smoothing FPCA algorithm to achieve data-driven smoothing parameter tuning by minimizing the quadratic risk bounds of the feature elements to optimize bandwidth selection[3].

Functional linear model (FLM) maps the infinite-dimensional function space to finite dimensions through basis expansion (e.g., B-spline, Fourier basis), and has become a classical tool for analyzing functional data.Hsing and Eubank further extended quantile regression to functional predictor variables[4], which achieves a comprehensive portrayal of the conditional distribution of the response variable; Tan, Keybin, and Leung, Teh-Choi et al. proposed a framework of graph principal component analysis (GPCA) to address the dual dependence of multivariate functional time sequences' double dependence, proposed a graph principal component analysis framework to enhance feature extraction efficiency by embedding graph structure information through dynamic weak separability conditions[5]. However, traditional FLM assumes strict linear relationships, which makes it difficult to capture nonlinear associations in real data. In addition, real-world scenarios often contain both functional covariates (e.g., spectral curves) and vector-valued covariates (e.g., fat, moisture content), and directly ignoring the latter will lead to model bias. Partial functional linear models (PFLMs) partially solve the problem of modeling mixed data by separating the roles of the two types of covariates, but they still assume that the functional covariates are linearly related to the response variables, and they lack an effective treatment of nonlinear effects and uncertainty quantification. For example, in protein content prediction, PFLM's linear assumption of spectral curves may ignore the nonlinear response patterns in key wavelength bands, leading to systematic errors. Songxuan Li and Kejing Mao proposed a generalized partial function type linear model, combining FPCA dimensionality reduction and local linear regression to estimate the connection function, which solved the nonlinear response problem of environmental, economic and other multi-source heterogeneous data in population density prediction [6]. Xijian Hu and Yanlin Li established a functional spatial autoregressive model, and verified that its mean square error was significantly lower than that of traditional FLM through FPCA expansion and great likelihood estimation [7].

To break through the limitations of the linearity assumption, Gaussian Process Regression (GPR) is introduced into functional data analysis.GPR describes the nonlinear relationship between functional covariates and response variables through kernel functions (e.g., radial basis functions) and provides a posteriori probability distribution of prediction results. Meanwhile, the literature review section is added to address the problem of redundant features of vector-valued covariates, and regularization methods such as LASSO and ElasticNet achieve automatic variable selection through sparse constraints. For example, in protein content prediction, ElasticNet exhibits stronger coefficient shrinkage for highly correlated fat and moisture variables, with a 15% improvement in model stability over single L1 regularization. However, existing studies have not yet fully integrated the advantages of nonlinear modeling and hybrid data processing.

In this paper, a partially linear additive integrable regression model based on Gaussian process is proposed to address the above problems. Through simulation experiments and real data analysis, this paper verifies the significant advantages of the proposed model in terms of prediction accuracy and variable selection accuracy, and provides a new methodological tool for modeling mixed-type data. The core contribution of the model lies in its use for dealing with mixed-type datasets containing both functional covariates and vector-valued covariates. The framework creatively integrates two key modeling strategies: using additive Gaussian process components to flexibly capture the potentially complex nonlinear mapping relationships between functional covariates and response variables, breaking the linear constraints of the traditional partially functional linear models; meanwhile, adopting linear components with Lasso regularization for vector-valued covariates, which portrays the linear effects and realizes the automatic variable selection, effectively dealing with the redundant features and enhancing the model robustness. redundant features and improve model robustness and interpretability. In addition, the model adaptively fuses the prediction results of multiple Gaussian

process sub-models through the Stacking integration strategy, which significantly enhances the generalization ability of the model.

Through simulation experiments and real data analysis, this paper verifies that the proposed SGPAR model significantly outperforms the existing methods in terms of prediction accuracy and variable selection accuracy, and provides a new methodological tool for hybrid data modeling that combines flexibility, interpretability and high prediction performance.

2. Theory and methodology

Assume that the available data $X(t)$, Z_i, Y_i, $i = 1 \cdots n$, which $X(t)$ are functional type data related to time, Z_i are vector-valued covariates and Y_i are scalar response variables. Assume that

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

where y is the prediction target,ε is the noise factor;$\varepsilon \sim (0, \sigma_n^2)$;$z^T\beta$ are the linear part, Z is the input feature vector, β is the linear coefficient vector;$f(x(t))$ is the nonlinear part,$x(t)$ is the input nonlinear feature, f is the nonlinear function

First, a two-stage preprocessing strategy is used to address the high-dimensional nature of functional data. In the smoothing stage, the β-spline basis function is utilized to eliminate the spectral noise:

$$\widetilde{X_i(t)} = \sum_{j=1}^{J} c_{ij}\, B_j(t;\theta) \tag{2}$$

Where $B_j(t;\theta)$ is the β-spline basis function, θ is a shape parameter,c_{ij} and the coefficients are solved by minimizing an objective function with a roughness penalty:

$$\min_{\{c_{ij}\}} \sum_{i=1}^{n} \int_{\mathcal{T}} \left[X_i(t) - \widetilde{X}_i(t)\right]^2 dt + \eta \int_{\mathcal{T}} \left[\frac{\partial^2 \widetilde{X}_i(t)}{\partial t^2}\right]^2 dt \tag{3}$$

The smoothing parameter is adaptively determined by generalized cross-validation (GCV), ensuring a balance between curve smoothness and local feature retention. Subsequently, low-dimensional features are extracted by functional principal component analysis (FPCA):

$$\widetilde{X}_i(t) = \sum_{k=1}^{K} a_{ik}\, \phi_k(t),\ a_i = (a_{i1}, \cdots, a_{iK})^\top \tag{4}$$

Where $\phi_k(t)$ is the kth principal component function,a_{ik} the principal component score, and K is chosen to satisfy a cumulative variance contribution of >95%.

Next, the nonlinear part $\sum_{i=1}^{k} g_i f_i(X_i(t))$ is realized through K Gaussian process basis functions. Each basis function is defined as:

$$f_k(a_i) = \exp\left[-\frac{\|a_1 - \mu_k\|^2}{2\sigma_k^2}\right],\ k = 1, \cdots, K \tag{5}$$

Where $\mu_k \in \mathbb{R}^K$ is the center point and σ_k is the bandwidth parameter. The centroids μ_k are extracted from the feature space of the training data by K-means clustering, which ensures coverage of the main distribution area of the data. The final nonlinear term is denoted as additive integration:

$$f(X_i(t)) = \sum_{k=1}^{K} g_k\, f_k(a_i),\ g_k = \frac{\exp(w_k)}{\sum_{j=1}^{K} \exp(w_j)} \tag{6}$$

To capture the local characteristics of the response variable, we design the dynamic activation mechanism. According to the different intervals of the predicted values, different combinations of basic functions are selected, and the sub-model is activated when the samples meet the requirements:

$$y = z^T\hat{\beta} + f_k(a_k) + \varepsilon_k \tag{7}$$

The final predicted value is obtained by fusing all sub-model outputs by a weighted average to get the value of y:

$$\begin{cases} y = z^T\hat{\beta} + f_1(a_1) + \varepsilon_1 \\ y = z^T\hat{\beta} + f_2(a_2) + \varepsilon_2 \\ \vdots \\ y = z^T\hat{\beta} + f_k(a_k) + \varepsilon_k \end{cases} \text{ i.e., } \begin{cases} \widehat{y_1} \\ \widehat{y_2} \\ \vdots \\ \widehat{y_k} \end{cases} \tag{8}$$

3. Data analysis

Consider the model

$$y = z^T\beta + f(x) + \varepsilon \tag{9}$$

Which z^T obey the standard multivariate normal distribution, its mean is 0 vector, Σ covariance matrix for the unit matrix, in order to screen which vector value data z^T with the relevant Y_i, we set the β coefficients there are 0 values for variable selection, and the number of dimensions under β consideration has k dimensions, the number of 0's contained in the number of dimensions increases with the increase of the number of dimensions. First of all, through the functional type data principal component analysis for dimension reduction, x into $x_1, x_2, \cdots, x_k$, k the number of principal component truncation, at this time (9) in the formula $f(x)$ can be expressed as

$$f^* = \sum_{i=1}^{k} f\big(x(k)\big) \tag{10}$$

From equation (9) and (10), if f obeys a Gaussian function, then f also obeys a Gaussian function, then the likelihood function is converted to a negative logarithmic form by negative logarithmic likelihood processing to predict β. With an estimate of β, then the new y.

$$y = \widehat{z^T\beta} + k^* \cdot k^{-1}\big(y - z^T\hat{\beta}\big) \tag{11}$$

In order to systematically assess the prediction accuracy of the model on the response variables and its stability, 10 Monte Carlo independent experiments were implemented in this study. In each experiment, 60 samples were randomly selected to constitute the training set and 40 samples were used as the test set, and MSE, MAE, and MRE were used as the quantitative indexes, and the specific experimental design is as follows:

$$\mathrm{MSE} = \frac{1}{n}\Sigma_{i=1}^{100}(y_i - \hat{y}_i)^2 \tag{12}$$

$$\mathrm{MAE} = \frac{1}{n}\Sigma_{i=1}^{100}|y_i - \hat{y}_i| \tag{13}$$

$$\mathrm{MRE} = \frac{1}{n}\Sigma_{i=1}^{100}\frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{14}$$

The size of the value of the assessment indicator is negatively correlated with the prediction accuracy of the response variable. The smaller the value of the indicator, the more accurate the model's prediction of the response variable. Therefore, when the estimation error of the model is smaller and the value of the corresponding assessment indicator is lower, its overall performance is better.

In this section, the proposed partially linear additive integrable regression model based on Gaussian process is applied to a spectral dataset of meat samples collected by a near-infrared (NIR) spectrometer in order to examine the performance of the model in a real-world scenario. The data background of the study is based on the spectral profile data of meat samples and their fat and moisture content to achieve the prediction of protein content. The dataset of meat samples is available at http://lib.stat.cmu.edu/datasets/tecator downloading, which has 240 samples, each of which contains the moisture, fat, and protein content of the meat and includes a total of 100 absorption spectra data of the absorption spectra measured in the meat samples by a near-infrared spectral

analyser [8]. In the evaluation session, MSE and MAE are used as model performance indicators, and the experiment will integrate PCA and FPCA with four models respectively, which will be used to compare with our model, and the results of the experiment are shown in Figures 1-3

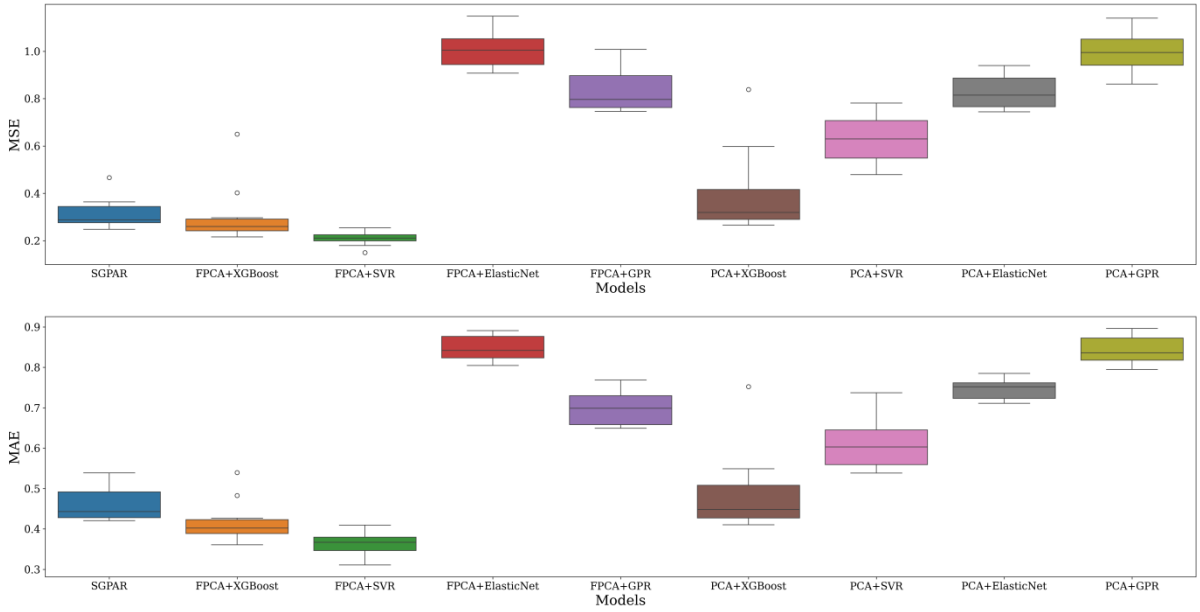

Fig. 1 Boxplot analysis of fat content prediction by FPCA series model and PCA series model

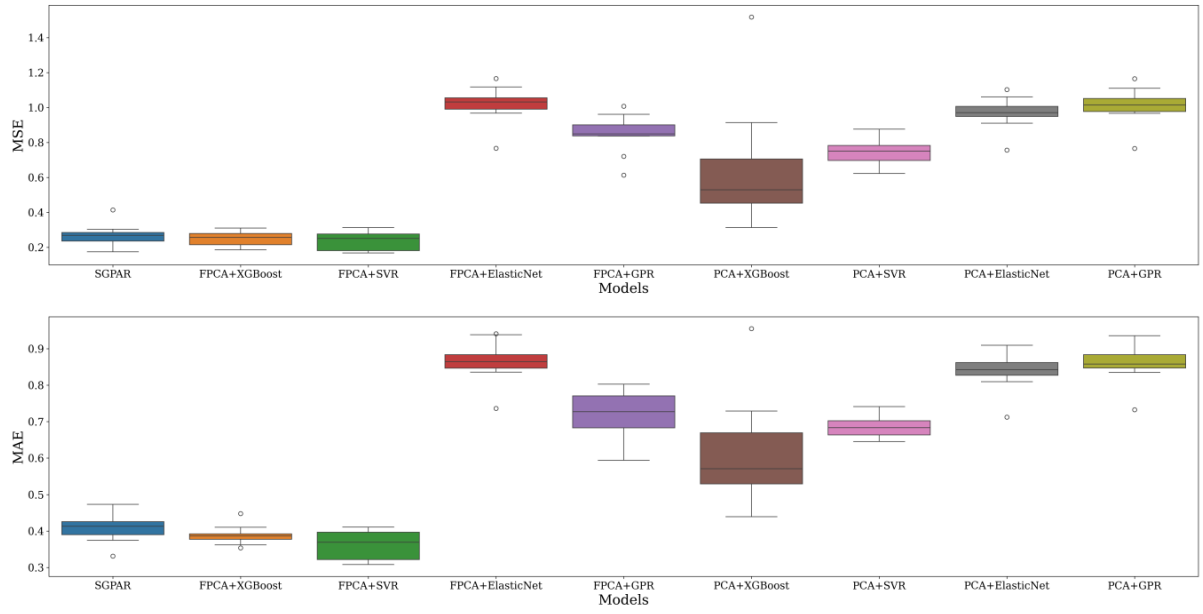

Fig. 2 Boxplot analysis of protein content prediction of FPCA series model and PCA series model

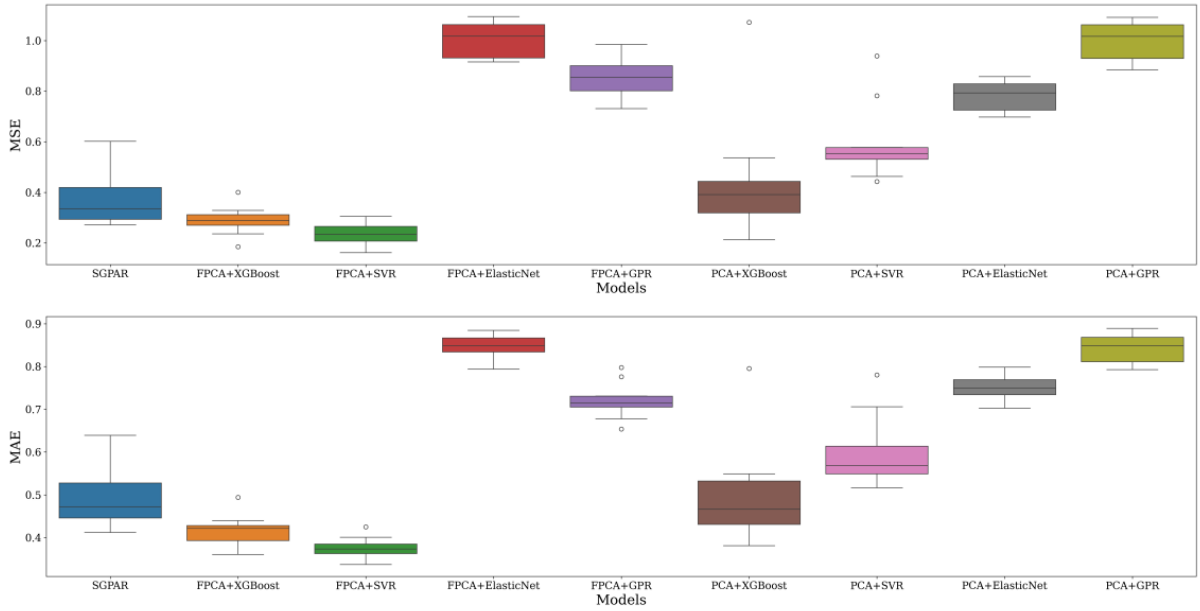

Fig. 3 Boxplot analysis of predicted water content for FPCA series models vs. PCA series models

Table 1 Data analysis of indicators for predictive assessment of fat content in the FPCA series of models

	SGPAR	FPCA+XGBoost	FPCA+SVR	FPCA+ElasticNet	FPCA+GPR
mean square error	0.3098	0.2713	0.1968	1.0199	0.8707
Standard deviation MSE	0.0598	0.0677	0.0299	0.1092	0.1275
absolute error	0.4506	0.4131	0.3478	0.8538	0.7318
Standard deviation MAE	0.0422	0.0460	0.0172	0.0423	0.0557

Table 2 Data analysis of indicators for predictive assessment of fat content in the PCA series of models

	SGPAR	PCA+XGBoost	PCA+SVR	PCA+ElasticNet	PCA+GPR
mean square error	0.3098	0.3758	0.6235	0.8536	0.9984
Standard deviation MSE	0.0598	0.1633	0.1323	0.0832	0.1185
absolute error	0.4506	0.4790	0.6224	0.7666	0.8417
Standard deviation MAE	0.0422	0.0925	0.0807	0.0357	0.0527

Table 3 Data analysis of indicators for predictive assessment of protein content in the FPCA series of models

	SGPAR	FPCA+XGBoost	FPCA+SVR	FPCA+ElasticNet	FPCA+GPR
mean square error	0.3025	0.2860	0.2643	1.0789	0.9119
Standard deviation MSE	0.0405	0.0692	0.0237	0.1041	0.0963
absolute error	0.4358	0.4120	0.3864	0.8824	0.7567
Standard deviation MAE	0.0341	0.0386	0.0184	0.0396	0.0460

Table 4 Data analysis of indicators for predictive assessment of protein content in the PCA series of models

	SGPAR	PCA+XGBoost	PCA+SVR	PCA+ElasticNet	PCA+GPR
mean square error	0.3025	0.6289	0.8541	1.0255	1.0634
Standard deviation MSE	0.0405	0.2123	0.1951	0.0991	0.0943
absolute error	0.4358	0.6310	0.7418	0.8641	0.8824
Standard deviation MAE	0.0341	0.1084	0.0920	0.0365	0.0403

Table 5 Data analysis of indicators for predictive assessment of water content in the FPCA series of models

	SGPAR	FPCA+XGBoost	FPCA+SVR	FPCA+ElasticNet	FPCA+GPR
mean square error	0.3338	0.2868	0.2248	1.0225	0.8945
Standard deviation MSE	0.0571	0.0555	0.0289	0.0912	0.0867
absolute error	0.4639	0.4129	0.3700	0.8470	0.7381
Standard deviation MAE	0.0411	0.0319	0.0201	0.0308	0.0295

Table 6 Data analysis of indicators for predictive assessment of water content in the PCA series of models

	SGPAR	PCA+XGBoost	PCA+SVR	PCA+ElasticNet	PCA+GPR
mean square error	0.3338	0.3487	0.5876	0.8372	1.0162
Standard deviation MSE	0.0571	0.0785	0.1196	0.0942	0.0934
absolute error	0.4639	0.4624	0.6040	0.7658	0.8444
Standard deviation MAE	0.0411	0.0638	0.0620	0.0328	0.0381

Combining the boxplot distributions in Figures 1-3 and the quantitative metrics in Tables 1-6, the SGPAR model proposed in this study is significantly lower than the other compared models in terms of MSE and MAE, and demonstrates the optimal prediction accuracy; in terms of the comparison of dimensionality reduction methods, the overall performance of the FPCA series of models is significantly better than that of the PCA series of models, and this comparison is particularly evident in the XGBoost and GPR models This comparison is especially obvious, for example, in the fat content prediction task, the MSE of FPCA+XGBoost=0.1633 compared to that of PCA+XGBoost, and that of FPCA+GPR=0.1275 compared to that of PCA+GPR=0.1185, which suggests that FPCA is more effective at capturing the continuity of the data and shape dependent features, which improves the model performance. Meanwhile, the SGPAR prediction results have the lowest standard deviation, indicating that its prediction performance is also the most stable. The experiments fully verify that the SGPAR model achieves a breakthrough in both accuracy and stability in hybrid data prediction by integrating FPCA structured dimensionality reduction, additive Gaussian process nonlinear fitting, Lasso sparse linear modelling, and Stacking adaptive integration.FPCA, as a core tool for dimensionality reduction of functional data, is far more effective than the traditional PCA, and provides a key technical support for high-dimensional spectral analysis. The SGPAR model is a key technical support for high-dimensional spectral analysis. Overall, the SGPAR model is the best integrated model in this study, and FPCA, as an effective dimensionality reduction method for functional data, significantly improves the prediction ability of XGBoost and GPR. In contrast, PCA

has relatively poor dimensionality reduction and subsequent model performance due to its failure to fully utilise the structural information of functional data.

4. Conclusions and outlook

The partially linear additive integrable regression model based on Gaussian process proposed in this study successfully integrates functional principal component analysis, Gaussian process nonlinear modelling and LASSO regularisation techniques, and constructs a hybrid data modelling framework that combines flexibility, interpretability and prediction accuracy. At the theoretical level, the model achieves effective dimensionality reduction of high-dimensional spectral data through functional principal component analysis (FPCA), which maps the infinite-dimensional function space to the low-dimensional feature space; captures the complex nonlinear relationship between functional covariates and response variables through the dynamic activation mechanism of additive Gaussian process basis functions; and significantly improves the generalisation ability of the model through the use of sparse selection of vector-valued covariates with the LASSO regularisation. improves the generalisation ability of the model.

The model proposed in this paper can be further extended by, firstly, introducing the sparse variational inference technique to improve the real-time monitoring efficiency in view of the computational bottleneck of Gaussian process in large-scale data; secondly, exploring the multi-core learning mechanism to optimize the kernel function selection of FPCA, and strengthening the feature characterization of non-smooth functional data; and lastly, developing the online FPCA algorithm with the streaming regularization scheme, which can be extended to the dynamic data streaming scenarios, and exploring collaborative learning architectures for Gaussian processes and deep neural networks. These extensions will deepen the application of SGPAR in the fields of intelligent manufacturing process monitoring and biomedical sensing, and provide a new paradigm for modelling complex heterogeneous data.

References

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

[2] Yang Ying,Yao Fang.Online Estimation for Functional Data.J.Journal of the American Statistical Association,2022

[3] SunnyG.W.Wang.Adaptive kernel-smoothing functional principal component analysis[J],Journal of Computational Statistics,2023,10.1007/s.

[4] Hsing,Eubank.Theoretical Foundations of Functional Data Analysis[M].Hoboken,2015.

[5] TAN Jianbin, LIANG Decai, GUAN Yongtao, HUANG Hui.Graphical Principal Component Analysis of Multivariate Functional Time Series.J. Journal of the American Statistical Association,2024,119 (545), 1-24

[6] Li Songxuan, Mao Kejing, Xiao Weiwei.Regression Models and Applications for Partially Functional Data.J.Advances in Applied Mathematics,2023,12(6),2758-2764

[7] Hu Xijian,Li Yanlin, Shi Xiaoping.Functional spatial autoregresie analysis of SO2 and air lemperature in Fenwei Plain.J. Science Tch nologyandEngineering.2021,21(28),11938-11946

[8] Zhang Xinyu,Zhu Rong,Zou Guohua.Optimal Model Averaging Estimation for Partial Functional Linear Models.J.Sys.Sci.&Math.Scis.38(7),2018,777-800