World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3078

PARTIAL NONLINEAR FUNCTIONAL REGRESSION MODEL: AN APPROACH VIA REPRODUCING KERNEL AND ENSEMBLE LEARNING

JiaYi Wang[1*], Yuan Chen[1], YanRan Liu[1], HaoDi Lv[2]
[1]*School of Economics and Management, Lanzhou University of Technology, Lanzhou 730050, Gansu, China.*
[2]*School of Foreign Languages, Lanzhou University of Technology, Lanzhou 730050, Gansu, China.*
[*]*Corresponding Author: JiaYi Wang*

Abstract: This study proposes a Partial Nonlinear Functional Regression Model (PNFLR) specifically designed to handle complex datasets where the Continuous Scalar Response Variable depends on a mixture of Vector-valued Covariates and Functional Covariates. The structural heterogeneity of these predictors is addressed by assuming a hybrid relationship: the vector components follow a standard Linear Association, whereas the functional inputs exhibit a complex Nonlinear Association with the response. To rigorously model this non-linearity within a Reproducing Kernel Hilbert Space (RKHS), the methodology departs from traditional single-kernel methods often characterized by rigid selection bias. Instead, the framework implements Model Averaging through an Ensemble Learning paradigm to facilitate the Adaptive Selection of kernel functions, thereby enhancing model flexibility. To ensure numerical stability and effective Regularization, a Truncated Approximation strategy is utilized. This process involves projecting the high-dimensional functional data onto a finite subspace via Functional Principal Component Basis Expansion, effectively mitigating overfitting risks while retaining essential structural information. By integrating kernel theory with ensemble mechanics, the PNFLR framework bridges the gap between theoretical function estimation and practical predictive modeling. Empirical evaluations on the Tecator dataset confirm that the architecture articulated herein yields superior Generalization Performance and lower error variance compared to conventional benchmark models across various prediction tasks, demonstrating robustness in real-world analytical scenarios.
Keywords: Functional Data Analysis (FDA); Functional regression; Reproducing kernel; Ensemble learning

1 INTRODUCTION

The explosive growth of data acquisition hardware has fundamentally altered the statistical landscape, inundating researchers with high-frequency functional datasets. This shift positions functional regression not merely as a theoretical exercise, but as the operational core of Functional Data Analysis (FDA). Consider the financial domain, where the objective is to map high-frequency exchange rate trajectories directly onto option implied volatility surfaces. Such modeling captures the intricate Nonlinear Association governing how volatility dynamics drive option pricing. However, treating these continuous structures merely as discrete High-dimensional Time Series is methodologically flawed. This reductionist approach discards essential smoothness properties and the strong Autocorrelation inherent in the data. Standard multivariate machine learning models therefore struggle, as they fail to account for the infinite-dimensional nature of the input. Developing a regression framework that explicitly respects this functional topology is thus mandatory.

Recent methodological advancements in functional regression have addressed increasingly specific structural challenges. Luo et al. integrated Functional Principal Component Basis Expansion with the Polya-Gamma Transformation to construct a Functional Cumulative Logistic Regression Model[1]. This framework employs a Gibbs Sampling Algorithm to implement precise Bayesian Estimation. Focusing on local geometric structures, Liu et al. derived kernel weights from curve distances to achieve locally weighted fitting of the Response Variable. Addressing data incompleteness[2]. Yang et al. proposed a Functional Nonparametric Quantile Regression Model robust to random missingness[3]. In geodetic applications, Tomohisa and Yukitoshi utilized ABIC basis expansion to estimate strain rate fields from GNSS datasets[4].

Empirical observation frequently indicates that the Response Variable is influenced jointly by Functional Covariates and scalar or Vector-valued Covariates. This hybrid dependency defines the structure of the Partial Functional Regression Model. In longitudinal medical studies, for instance, prognostic outcomes depend on the interaction between continuous physiological monitoring curves and static variables such as age and comorbidities. Similarly, in environmental science, Air Quality Indices reflect both high-frequency meteorological fluctuations and discrete socio-economic determinants like industrial output. To address these mixed inputs, Li et al. combined Functional Principal Component Analysis (FPCA) with Maximum Likelihood Estimation (MLE) to estimate coefficients for Vector-valued Covariates[5]. They further utilized Local Linear Regression to approximate the link function, facilitating adaptive modeling. Ling et al. developed a k-Nearest Neighbor Estimation (k-NN) method to integrate functional linear components with nonlinear vector segments[6]. Regarding statistical diagnostics, Wen examined the Homogeneity of Variance Test within this framework[7]. Huang et al. focused on estimation paradigms under measurement error

conditions, while Zhu et al. investigated Model Averaging mechanics[8,9]. Additionally, Liu et al. proposed a Composite Quantile Regression framework for varying-coefficient models[10]. However, the majority of extant research relies on the restrictive assumption of a strict Linear Association between the Functional Covariate and the Response Variable. This premise often contradicts empirical reality, where Nonlinear Association is pervasive rather than anomalous. Consequently, relaxing this linearity constraint represents a necessary theoretical advancement.

This study investigates the Nonlinear Association between Functional Covariates and the Response Variable. We model this nonlinearity using Reproducing Kernel theory. Since the choice of kernel function is critical for Generalization Performance, relying on a single kernel carries significant risk. Therefore, we avoid singular model selection in favor of Model Averaging. By employing Ensemble Learning, we achieve the Adaptive Selection of optimal kernels. The procedure is as follows: first, we construct a set of base models, each using a different kernel function. These models are trained in parallel, and their outputs are integrated into the ensemble framework. This aggregation produces a final prediction with enhanced Robustness. By balancing the trade-off between bias and variance, this approach effectively mitigates both Overfitting and Underfitting. Empirical results confirm that this architecture consistently outperforms conventional benchmark methods.

2 THEORY AND METHODOLOGY

Formally, the Partial Nonlinear Functional Regression Model (PNFLR) advanced herein is defined as:

$$Y = \mathrm{z}^T \beta + f\left(x\left(t\right)\right) + \varepsilon \tag{1}$$

Where Y represents the Response Variable, and z denotes the Vector-valued Covariate. The observed spectral curve is represented by $x(t)$, which is assumed to reside within a Reproducing Kernel Hilbert Space (RKHS). Both β and f are unknown parameters requiring estimation. Specifically, β constitutes the linear component of the model and is estimated via Penalized Least Squares. Because $x(t)$ belongs to the RKHS, the nonlinear function $f(x)$ can be expressed using a Reproducing Kernel as:

$$f\left(x\right) = \left\{ \sum_{i=1}^{n} \partial_i K\left(x_i\left(t\right), x\left(t\right)\right) \mid \partial_i \in R \right\} \tag{2}$$

The kernel function $K\left(x, x_i\right)$ is instantiated as a Gaussian Kernel Function, defined as follows:

$$K\left(x_1\left(t\right), x_2\left(t\right)\right) = e^{\frac{-\|x_1(t)-x_2(t)\|^2_H}{2\delta^2}} \tag{3}$$

Within this Hilbert space H, the distance metric between any two functional elements, denoted as $x_1\left(t\right), x_2\left(t\right)$, is formally defined as:

$$\| x_1\left(t\right) - x_2\left(t\right) \|^2 = \int_{t\in J} \left(x_1\left(t\right) - x_2\left(t\right)\right)^2 dt \,, \tag{4}$$

Subsequently, by subjecting the aforementioned expression to a Basis Expansion procedure.

$$x_1\left(t\right) = \sum_{j=1}^{J} a_{1j}\phi_j\left(t\right), x_2\left(t\right) = \sum_{j=1}^{J} a_{2j}\phi_j\left(t\right), \tag{5}$$

The Formulation is resolved into:

$$\int_{t\in J} \left(x_1\left(t\right) - x_2\left(t\right)\right)^2 dt = \left(a_1 - a_2\right)^T \int \Phi * \Phi^T dt \left(a_1 - a_2\right) \tag{6}$$

We define the coefficient vectors $a_1 = (a_{11}, a_{12}, \cdots, a_{1j})$ and $a_2 = (a_{21}, a_{22}, \cdots, a_{2j})$, alongside the basis function vector $\Phi = (\phi_1(t), \phi_2(t), \cdots, \phi_J(t))$. Consequently, the following structural relationships are established for the model:

$$\begin{aligned} Y &= z^T \beta + f\left(x\left(t\right)\right) + \varepsilon \\ &= z^T \beta + \sum_{i=1}^{n} \alpha_i K\left(x_i\left(t\right), x\left(t\right)\right) + \varepsilon \\ &= z^T \beta + \left(\alpha^T \bullet \Phi\right)^T + \varepsilon \\ &= \left(z, \Phi\right)^T \begin{pmatrix} \beta \\ \alpha \end{pmatrix} + \varepsilon \end{aligned} \tag{7}$$

Here, the weight vector alpha is defined as $\alpha = (\alpha_1, \cdots \alpha_n)^T$, and Phi represents the kernel transformation vector $\Phi = (K(x_1(t), x\ (t)), \cdots, K(x_n(t), x(t)))^T$. By substituting the n sets of observational data into the previous formulation, we obtain the following system for i = 1, 2, ..., n: $Y_i = (z_i, \Phi(x_i(t)))^T (\alpha^{\beta}) + \varepsilon_i, i = 1, 2, \cdots, n$.

To estimate the unknown parameters, we employ Penalized Least Squares. This optimization approach allows us to solve for both the linear coefficients and the nonlinear expansion weights simultaneously. Specifically, the objective function is constructed as follows:

$$\tilde{\alpha},\ \tilde{\beta} = \arg\min_{\alpha,\beta} \frac{1}{n} \sum_{i}^{n} (Y_i - (z_i, \Phi(x_i(t))(\alpha^{\beta}))^2 + \lambda \alpha^T K \alpha + \lambda_2 \|\beta\| \tag{8}$$

Solving this minimization problem yields the final Parameter Estimation for the model.

Next, noting that the choice of kernel function K directly affects model performance and that a single kernel often fails to adapt to complex structural patterns in the data—leading to overfitting or underfitting—this paper adopts an ensemble learning framework. We select M different kernel functions $\{K_m\}_{m=1}^{M}$,and construct corresponding regularized regression estimators to obtain a set of heterogeneous base models. The predictions of these base models on the training data

$$\hat{\mathbf{Y}}_m = \left(\hat{f}_m(x_1, z_1), \ldots, \hat{f}_m(x_n, z_n)\right)^{\top} \in \mathbb{R}^n \tag{9}$$

are combined into the feature matrix

$$\mathbf{H} = [\hat{\mathbf{Y}}_1, \hat{\mathbf{Y}}_2, \ldots, \hat{\mathbf{Y}}_M] \in \mathbb{R}^{n \times M} \tag{10}$$

which is used as input to train a meta-regression model $g : \mathbb{R}^M \to \mathbb{R}$. Finally, the prediction for a new sample *(x,z) is given in ensemble form as*

$$\hat{Y} = g\left(\hat{f}_1(x, z), \hat{f}_2(x, z), \ldots, \hat{f}_M(x, z)\right) \tag{11}$$

By fusing heterogeneous information from multiple kernel mappings, this approach effectively improves the model's robustness and generalization performance.

3 EMPIRICAL ANALYSIS

3.1 Data Sources and Experimental Configuration

We analyze the meat processing dataset obtained from the Tecator archive (http://lib.stat.cmu.edu/datasets/tecator). The data comprises $n = 240$ samples, each documenting moisture, fat, and protein content. Additionally, Near-Infrared (NIR) absorbance spectra are recorded across the 850–1050 nm range. Given the 2 nm sampling interval, each spectrum consists of 100 discrete points. The goal is to predict protein content using these spectral curves alongside the scalar physicochemical measurements.

A preliminary visual inspection of the dataset is presented in Figure 1. The left panel delineates the multi-curve trend, capturing the global morphology, fluctuation modes, and dynamic range of the samples. Observably, the curves exhibit a high degree of morphological consistency, attesting to the superior quality of the dataset, which is devoid of conspicuous outliers. Complementarily, the right panel displays a correlation heatmap illustrating the interrelationships among distinct features. The visualization reveals an intensely high correlation structure, strongly signaling the presence of severe multicollinearity and data redundancy. Under such conditions, the application of conventional linear regression models would likely result in unstable coefficient estimates and diminished generalization capabilities. Consequently, to mitigate these methodological risks, this study adopts a Functional Data Regression modeling approach.

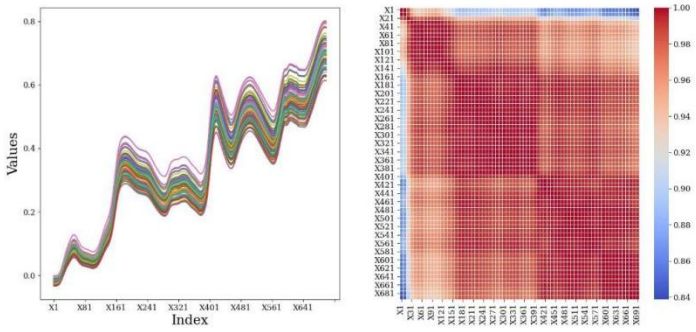

Figure 1 Visual Characterization of the Meat Data Structure

We designed three experimental configurations to rigorously evaluate the model. In the first scenario, Protein serves as the predictor for Fat and Water. The second scenario uses Fat to predict Protein and Water, while the third employs

Water to predict Fat and Protein. The analysis includes 216 samples. To capture various data structures, we utilize Gaussian, Polynomial, Linear, and Laplacian kernels. A Random Forest model then acts as an ensemble meta-learner to integrate these kernel outputs. Validation follows a Monte Carlo Cross-Validation protocol. We randomly partition the dataset into training and testing sets over repeated iterations. We aggregate the resulting error metrics to calculate the Mean and Standard Deviation (SD). A lower Mean signifies better accuracy, while a smaller SD indicates improved stability. Consequently, we rely on the Mean Squared Error (MSE) and its deviation to quantify performance.

3.2 Comparative Experimental Results

Tables 1-3 summarize the Mean and Standard Deviation (SD) for the three sub-tasks, while Figures 2-4 display the corresponding boxplots. The results demonstrate that the PNFLR method consistently achieves the lowest Mean and SD. The boxplots further corroborate this stability, showing that PNFLR maintains the narrowest interquartile range compared to other methods. This performance advantage stems from two key factors. First, we strictly treat multi-dimensional time series as Functional Data rather than discrete points. Second, the model explicitly captures the intrinsic Nonlinear Association between functional covariates and predictors. Furthermore, by integrating Ensemble Learning, the framework avoids the limitations of single-kernel approaches, thereby significantly enhancing both Generalization Performance and robustness.

Table 1 Experimental Results Utilizing Protein as the Predictor Variable

Methods	Mean	Standard Deviation
NPFLR	0.1485	0.0352
XGBoost	0.4799	0.0833
RF	0.4834	0.0494
SVR	0.6967	0.0975
ElasticNet	0.9942	0.0996
GPR	0.7775	0.1245

Table 2 Experimental Results Utilizing Water as the Predictor Variable

Methods	Mean	Standard Deviation
NPFLR	0.019	0.006
XGBoost	0.3415	0.0978
RF	0.3353	0.0828
SVR	0.51	0.0993
ElasticNet	0.9976	0.1202
GPR	0.8522	0.1422

Table 3 Experimental Results Utilizing Fat as the Predictor Variable

Methods	Mean	Standard Deviation
NPFLR	0.0224	0.0051
XGBoost	0.3881	0.0653
RF	0.3715	0.0845
SVR	0.5372	0.0722
ElasticNet	0.9713	0.0848
GPR	0.8144	0.1012

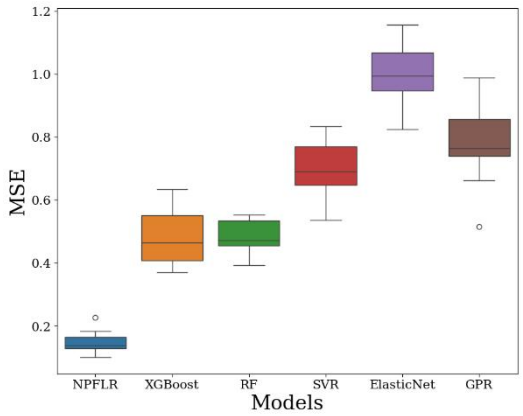

Figure 2 Visualization of Prediction Outcomes with Protein Content as the Response Variable

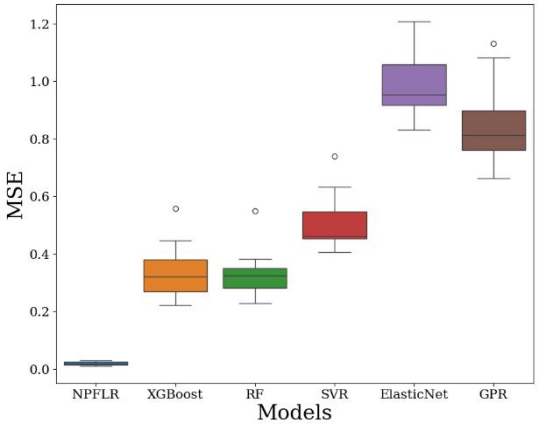

Figure 3 Visualization of Prediction Outcomes with Water Content as the Response Variable

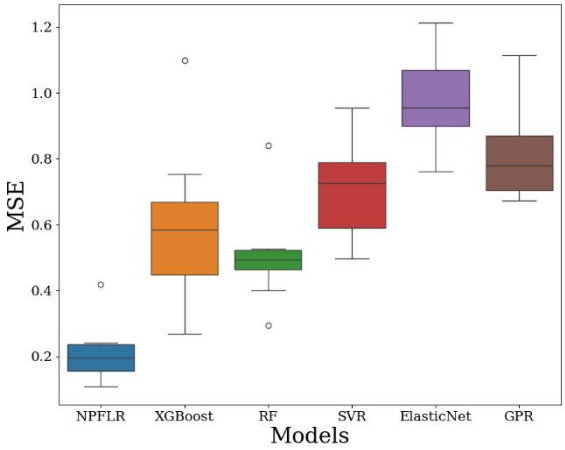

Figure 4 Visualization of Prediction Outcomes with Fat Content as the Response Variable

4 CONCLUSION

The Partial Nonlinear Functional Regression Model (PNFLR) established in this study successfully integrates Reproducing Kernel theory with Ensemble Learning to address complex data structures comprising both functional and vector-valued covariates. By accommodating the nonlinear dependency between the response and functional predictors, the model utilizes kernel functions to rigorously fit intrinsic data patterns. Crucially, the framework moves beyond the limitations of single-kernel selection by implementing an Ensemble Learning strategy. This Model Averaging approach not only yields robust predictions but also effectively balances the trade-off between bias and variance, thereby enhancing the model's stability and Generalization Performance. Empirical evidence from real-world data analysis substantiates the efficacy of this architecture, demonstrating predictive accuracy superior to that of prevailing benchmark techniques.

Reflecting on these findings highlights several distinct avenues for future inquiry. One primary challenge lies in the functional data processing stage, specifically regarding the truncated approximation via Principal Component Analysis (PCA). The theoretical determination of an optimal truncation threshold remains a non-trivial issue; future

investigations will focus on resolving this by incorporating cross-validation techniques or information criteria to conduct quantitative research and establish rigorous standards. Furthermore, the current assumption that vector-valued covariates maintain a strictly linear relationship with the response may be relaxed. Subsequent work should explore more sophisticated fitting methodologies to capture potential nonlinearities within this vector segment. Lastly, the scope of the ensemble framework warrants expansion. Designing comprehensive comparative protocols to benchmark alternative Ensemble Learning architectures constitutes a vital step toward further methodological refinement.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Luo Youxi, Deng Nan, Hu Chaozhu, et al. Research and application of functional cumulative logistic regression model. Journal of Central China Normal University (Natural Sciences), 2023, 57(2): 185-194.
[2] Liu Xinyang, Li Xiuying, Geng Fazhan. A nonparametric regression method for functional data based on kernel function. Journal of Changshu Institute of Technology, 2025, 39(2): 103-106.
[3] Yang Jintao, Ling Nengxiang. Estimation of functional nonparametric quantile regression model with randomly censored response variables. Journal of Hefei University of Technology (Natural Science), 2023, 46(5): 709-712.
[4] Tomohisa O, Yukitoshi F, Takuya N. Consistent estimation of strain-rate fields from GNSS velocity data using basis function expansion with ABIC. Earth, Planets and Space, 2021, 73(1).
[5] Li Songxuan, Mao Kejing, Xiao Weiwei. Regression models and applications for partially functional data. Advances in Applied Mathematics, 2023, 12(6): 2758-2764.
[6] Wen Liyu. Homogeneity test of variance for partially functional linear regression model. Beijing: Beijing University of Technology, 2022.
[7] Ling N, Aneiros G, Vieu P. kNN estimation in functional partial linear modeling. Statistical Papers, 2022, 61, 423-444.
[8] Huang Jiewu, Wang Linjie, Chen Xingyue, et al. Research on estimation of partial functional linear models with measurement errors. Journal of Tonghua Normal University, 2025, 46(6): 26-32.
[9] Zhu Rong, Zou Guohua, Zhang Xinyu. Model averaging method for partial functional linear models. Journal of Systems Science and Mathematical Sciences, 2018, 38(7): 24.
[10] Liu Yanxia, Wang Zhihao, Tian Maozai. Composite quantile regression estimation of varying coefficient partially functional linear models. Journal of Mathematical Statistics and Management, 2025(2): 1-13.