

Functional Bayesian Model Average Regression Based on Derivative Information

Zibing Wang *

School of economics, Henan University, Kaifeng, China, 475001

* Corresponding Author Email: neve_w@163.com

Abstract. Functional data regression models based on basis function expansion have been widely used in economic, financial, and environmental fields. However, determining the appropriate type and quantity of basis functions, such as KL expansion and B-spline, has remained a challenge to overcome. Additionally, when two curves are similar, constructing an effective regression model can prove to be difficult. On top of that, this paper suggests a functional Bayesian model average regression model, which is based on the KL expansion of the derivative function. On the one hand, the method uses the first and second-order derivative function information as predictors to capture the curve's fluctuation trend more comprehensively. On the other hand, the method applies model averaging to dynamically determine the appropriate number of basis functions to use, which can effectively alleviate the overfitting and underfitting problems of the model. The experiments and corresponding data analysis demonstrate that the proposed method has higher prediction accuracy and robustness than the comparative methods.

Keywords: Functional regression, Bayesian model averaging, The derivative function characterization.

1. Introduction

With the rapid development and progress of technology, data has become the core driving force of modern society. In many practical applications, data often exist in the form of images or curves, which have distinctive functional characteristics and are called functional data. Functional data analysis, as an emerging research field, aims to extract valuable information from these complex functional data to support scientific research, engineering design, business decision-making and other fields.[1]

Functional data regression modeling is a significant approach in the field of functional data analysis [2]. One of the most widely discussed problems is regression modeling with functional data as covariates and scalars as response variables. For this type of problem, researchers have proposed a variety of estimation methods. One is to directly select the basis functions to base expand the covariates and slope functions; the other is to construct the basis functions based on the data information, such as the functional principal component basis functions. Reiss, P.T et al. (2017) used a functional linear model to characterize the relationship between the functional covariates and the scalar response variable [3]. Cardot et al. (2003)[4], Hall and Hooker (2016) [5] use the smooth spline technique and functional principal component analysis respectively to estimate the slope function and investigate the asymptotic properties of the estimates. Kong et al. (2016a)[6]; Su et al. (2017) [7] and Zhou et al.(2023) [8] investigated the problem of hypothesis testing about the slope function in functional type regression models using FPCA, where the truncation is in the form of the cumulative contribution rate. For more information on functional data regression modeling see DING H(2018)[9]. However, there are some limitations to each of these methods. For example, direct selection of basis functions for base expansion still yields infinite dimensional parameters, whereas samples are typically finite dimensional, and thus truncation of the base expanded model is required to ensure that model parameters can be estimated. This base expansion method is independent of the data characteristics and does not guarantee that the retained basis functions capture the main features of the functional-type data. On the other hand, using the approximation method of functional principal components to construct regression models tends to ignore the key information of functional data, for

example, the principal component function that may be closely related to the response variable is in the latter part of the eigenvalue series, and if the cumulative contribution truncation method is used, this part of the information will be lost, which will make the prediction of the regression model inaccurate.

To overcome these limitations and retain the key information of functional data, this paper proposes a functional type regression model based on the information of curve derivatives. For one thing, the proposed method extracts the information of first-order derivatives and second-order derivatives of the function-type data. This method can effectively retain the key information of the function-type data, thus improving the performance of the regression model. For another, the Bayesian model averaging method is introduced to solve the problems of overfitting or underfitting of the predictive model caused by improper selection of principal components after using KL expansion. The experiment suggests that the proposed method has a smaller prediction error and higher robustness than the cumulative contribution ratio method and the model selection method such as LASSO.

2. Theory and methodology

2.1. Functional Principal Component Analysis

Functional data refers to data that is represented in the form of a function, where the entire function is considered as a singular unit of data. However, in reality, the values obtained from observations are often discrete points, and there may be errors in the observations. Therefore, it is necessary to reconstruct the actual functions that are hidden behind the observation data.

Suppose that $X_i(t)$ ($i=1,2,\dots,n$) is the object of function under study, $y_{i1}, y_{i2}, \dots, y_{iT_i}$ are the T_i observations of function $X_i(t)$, ε_{ij} ($j = 1, 2, \dots, T_i$) is the observation error of the j -th observation value of object, which leads to equation(1):

$$y_{ij} = x_i(t_j) + \varepsilon_{ij} \quad (1)$$

for the estimation of $X_i(t)$ expand it over a set of basis functions $\Phi(t) = \{\phi_1(t), \phi_2(t), \dots, \phi_k(t)\}'$:

$$X_i(t) = \sum_{k=1}^K \phi_k(t) c_k^i \quad (2)$$

let $\mathbf{c}_i = (c_{i1}, c_{i2}, \dots, c_{ik})'$, represent Eq. (2) in the matrix form:

$$x_i(t) = \Phi(t)' \mathbf{c}_i \quad (3)$$

substituting Eq.(2) into Eq. (1), the base expansion coefficients are estimated according to the least square rule:

$$\mathbf{c}_i = \operatorname{argmin} \sum_{j=1}^{T_i} (y_{ij} - \sum_{k=1}^K \phi_k(t) c_k^i) \quad (4)$$

similarly, the first-order and second-order derivative of $x_i(t)$ are estimated:

$$x_i'(t) = \Phi(t)' \mathbf{a}_i \quad (5)$$

$$x_i''(t) = \Phi(t)' \mathbf{b}_i \quad (6)$$

Where the $\mathbf{a}_i$ and $\mathbf{b}_i$ are base-expanded coefficient vectors of $x_i'(t)$ and $x_i''(t)$ respectively. After smoothing the functions by basis expansion, the paper extracts K-dimensional vectors to represent the original samples. However, considering the text utilizing both first-order and second-order derivative information in the clustering distance, 2k-dimension features are extracted for each sample. In order to improve computing efficiency and reduce data noise, the text applies functional principal component analysis to reconstruct the samples. To start with, let $N = n - 1$, the sample mean function and sample covariance function are used to estimate the mean function $\mu(t)$ and covariance function $V(s, t)$:

$$\hat{\mu}(t) = \frac{1}{N} \sum_{i=1}^n x_i(t) \quad (7)$$

$$\hat{V}(s, t) = \frac{1}{N} \sum_{i=1}^n (x_i(s) - \hat{\mu}(s))(x_i(t) - \hat{\mu}(t)) \quad (8)$$

The covariance function is then spectrally analyzed to obtain the characteristic equation:

$$\int_0^T f_j(t) V(s, t) dt = \sum_{j \geq 1} \lambda_j f_j(s) \quad (9)$$

Where $\lambda_1 \geq \lambda_2 \geq \dots$, and when $j=j'$ $\int_0^T f_j(t) f_{j'}(t) dt=1$; when $j \neq j'$, $\int_0^T f_j(t) f_{j'}(t) dt=0$. Besides, $f_j(t)$ denotes eigenfunction and λ_j denotes eigenvalue. The basis function expansion of the characteristic function is given as $f(t) = \sum_{k=1}^K \phi_k(t) b_k$ i.e. $f(t) = \mathbf{b} \Phi(t)$. where $\mathbf{b} = (b_1, b_2, \dots, b_K)$ is the parameter to be estimated. Meanwhile, substituting Eq.(3) into Eq. (7) yields:

$$\hat{V}(s, t) = N^{-1} X' X = N^{-1} \Phi(s)' \mathbf{c}' \mathbf{c} \Phi(t) \quad (10)$$

replace $V(s, t)$ by $\hat{V}(s, t)$, and substitute the above two equations into the characteristic equation:

$$\int_0^T f_j(t) \hat{V}(s, t) dt = \int N^{-1} \Phi(s)' \mathbf{c}' \mathbf{c} \Phi(t) \Phi(t)' \mathbf{b} dt = N^{-1} \Phi(s)' \mathbf{c}' \mathbf{c} W \mathbf{b} \quad (11)$$

$$\sum_{j \geq 1} \lambda_j f_j(s) = \boldsymbol{\rho} \mathbf{b} \Phi(s)' \quad (12)$$

where $W = \int \Phi(t) \Phi(t)' dt$, $\boldsymbol{\rho} = (\lambda_1, \lambda_2, \dots)$. Organize the above equations and Eq.(11) can be simplified as:

$$N^{-1} \mathbf{c}' \mathbf{c} W \mathbf{b} = \boldsymbol{\rho} \mathbf{b} \quad (13)$$

which satisfies $\|f_j\| = 1$, i.e. $\mathbf{b}' W \mathbf{b} = 1$. If $\{f_1, f_2, \dots, f_L\}$ is a set of orthogonal basis functions, then by the $K - L$ Expansion Theorem, $x_i(t)$ can be expanded as:

$$x_i(t) \approx \hat{x}_i(t) = \mu(t) + \sum_{m=1}^L a_{im} f_m(t), L \leq K \quad (14)$$

where a_{im} stands for the main component scores:

$$a_{im} = \int x_i(s) f_m(s) ds \quad (15)$$

the global approximation criterion is to minimize the following objective function:

$$PCASSE = \sum_{i=1}^N \|x_i - \hat{x}_i\|^2 = \int_0^T [x(t) - \hat{x}(t)]^2 dt \quad (16)$$

similarly, the principal component expansion of $x_i'(t)$ and $x_i''(t)$ can be obtained as follows:

$$x_i'(t) \approx \hat{x}_i'(t) = \mu'(t) + \sum_{m=1}^L a'_{im} f'_m(t), L \leq K \quad (17)$$

$$x_i''(t) \approx \hat{x}_i''(t) = \mu''(t) + \sum_{m=1}^L a''_{im} f''_m(t), L \leq K \quad (18)$$

their eigenvalues are respectively $\boldsymbol{\rho}' = (\lambda'_1, \lambda'_2, \dots)$, $\boldsymbol{\rho}'' = (\lambda''_1, \lambda''_2, \dots)$ and their principal component score vectors are $\mathbf{a}'_{im} = (a'_{i1}, a'_{i2}, \dots, a'_{iL})'$; $\mathbf{a}''_{im} = (a''_{i1}, a''_{i2}, \dots, a''_{iL})'$. So far, each sample is downsampled from 2K features to 2L.

2.2. Bayesian model mean regression based on derivative information

Based on the idea of simultaneously incorporating the first-order derivative function and second-order derivative function information of function-type data into regression, this paper uses the first-order derivative function and second-order derivative function principal component scores of function-type data to be put into the vector $\mathbf{m}_k$. Then divide by $\boldsymbol{\lambda} = [\boldsymbol{\rho}', \boldsymbol{\rho}'']$ for normalization, and finally denote by $\mathbf{m}_k$ the information in the k th dimension of the KL unfolding of $X(t)$:

$$\mathbf{m}_k = \frac{[a'_{k'}, a''_{k'}]}{\lambda_k} \quad (19)$$

Considering that the degree of explanation of $X(t)$ by the vectors of each obtained after KL expansion is unknown, this paper adopts the Bayesian model averaging method to construct the regression model. The Bayesian model averaging method can reduce the overfitting and underfitting problems of a single model by averaging multiple models, and at the same time, it can reduce the uncertainty due to the selection of a single best model. As a result, when Bayesian averaging is applied on the regression model of $\mathbf{m}_k$ to y , the method balances the effects of different variables retained by the principal component analysis on $X(t)$, and solves the problem of overfitting or underfitting to a certain extent due to the fact that the degree of explanation of the KL -truncated variables for the function-type data is unknown. The weight selection criterion w_k is:

$$w_k = \frac{\exp(\frac{-BIC_k}{2})}{\sum_L \exp(\frac{-BIC_k}{2})} \quad (20)$$

k represents the k -th model, $BIC_k = -2\log\ell_k + l_k\log n$, ℓ_k and l_k are the great likelihood function and the number of unknown parameters of model k , respectively. A total of L principal components were retained in the upper stage, and the basic model of L principal components is:

$$y = \beta_k \mathbf{m}_k + \varepsilon_k \quad (21)$$

β_k is the regression coefficient of the k -th basic model, ε_k is the error term for the k -th fundamental model. According to OLS principle, estimate β_k :

$$\beta_k = \min \|y - \beta_k \mathbf{m}_k\|^2 \quad (22)$$

this leads to the regression model in this paper as:

$$y = \sum_L w_k \beta_k \mathbf{m}_k + \varepsilon \quad (23)$$

3. Data analysis

3.1. Experimental setup and data sources

In this paper, the meat dataset (downloaded from <http://lib.stat.cmu.edu/datasets/tecator>) and the corn dataset (downloaded from <http://www.eigenvector.com/data/Corn/index.html>) were chosen to conduct experiments in R. The meat dataset consists of 240 samples containing absorption spectra obtained by a near infrared (NIR) spectrum analyzer testing meat in the 850-1050 nanometer (nm) range, with a total of 100 absorption spectra recorded at intervals of two bands. The data set also included moisture, fat and protein content for each sample.

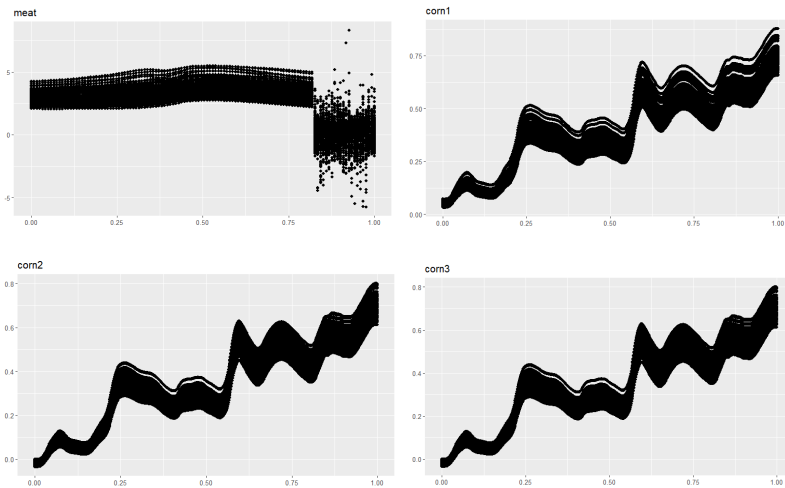


Figure 1 samples of the dataset

The maize dataset contains 80 samples with measurements from three different spectrometers, each measuring a total of 700 absorption spectra in the band 1100-2498 nm at 2 nm intervals. The

dataset similarly recorded the oil, protein moisture and starch content of the corn samples. Pairing the functional data in the dataset with the corresponding scalar data two by two, the two datasets yielded a total of 15 sets of data. Figure 1 illustrates four sets of functional type data from the experiment.

The experiments were designed to use the DFMA algorithm in this paper and four other comparative algorithms to predict the corresponding moisture, fat, protein, or starch content of meat or corn from their absorption spectrum data. The experiments were repeated 100 times for each of the 15 data sets in the dataset and the performance of the comparison algorithms was evaluated using the metrics in the following page. Meanwhile, this paper divides the dataset into training set and testing set, and estimates the model parameters through the training set and predicts the data through the test set to evaluate the model performance. The specific information is shown in Table 1 below:

Table 1 data set information

Data set	background	Training sample	Testing sample	dimension	X	y
Meat data	Infrared absorption spectral data of meat, as well as moisture, oil and starch content data.	40	40	100	1	3
Corn data	Infrared absorption spectral data of corn, as well as water, oil and starch content data.	140	100	700	3	4

In this paper, the mean and standard deviation of the mean square error between the predicted value and the true value of the model after 100 experiments are used as the evaluation index. The specific formula is as follows:

$$\mu_n = (\hat{y}_n - y_n)^2 \quad (24)$$

$$mean = \frac{\sum_{n=1}^{100} \mu_n}{100} \quad (25)$$

$$sd = \sqrt{\frac{\sum_n (\mu_n - mean)^2}{100}} \quad (26)$$

where μ_n stands for the mean square error between the predicted value and the true value in the $n - th$ experiment, $mean$ denotes the mean value of μ_n of 100 experiments, sd denotes the standard deviation of μ . $mean$ and sd are indicators in this paper for evaluation. (The metrics for evaluating the algorithms in this paper are denoted directly by $mean$ and sd in the following page)

3.2. Comparative results

In this paper, we compare the prediction results of the model on two datasets with those on four models: $fpca$ -OLS^[2], $bspline$ -OLS^[9], pca -OLS^[10], $fpca$ -LASSO^{[11][12]}. To simplify the representation, the absorption spectrum data in the meat dataset is denoted as MEAT, and the three absorption spectrum datasets in the corn dataset are denoted as CORN1, CORN2, and CORN3. The results of the comparison are shown in Table 2 below.

Table 2 results of experiment

X-y	DFMA		bspline-OLS		pca-OLS		fpca-OLS		fpca-LASSO	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Meat-Protein	1.0732	0.5825	10.1881	1.2583	8.5895	0.9936	11.1855	1.3806	10.5280	1.5103
Meat-Fat	12.018	5.8063	152.923	24.537	135.538	19.217	174.184	27.115	150.641	24.330
Meat-Moisture	8.5597	2.5645	86.4074	12.470	77.1427	10.318	98.6674	14.472	82.2550	12.832
Corn1-Moisture	0.0004	0.0001	0.1050	0.0168	0.11186	0.1705	0.1262	0.0196	0.0998	0.0165
Corn1-Protein	0.0080	0.0023	0.0319	0.0054	0.0258	0.0037	0.0321	0.0049	0.0323	0.0054
Corn1-Fat	0.0295	0.0010	0.2368	0.0309	0.2048	0.0280	0.2388	0.0319	0.2336	0.0304
Corn1-Starch	0.1069	0.0418	0.69219	0.1028	0.6477	0.0992	0.6857	0.1040	0.6898	0.1041
Corn2-Moisture	0.0321	0.0122	0.1082	0.0157	0.1111	0.0161	0.1261	0.1842	0.1001	0.0149
Corn2-Protein	0.0121	0.0035	0.0299	0.0050	0.0254	0.0037	0.0306	0.0049	0.0301	0.0053
Corn2-Fat	0.0348	0.0109	0.2461	0.0307	0.2405	0.0295	0.2543	0.3197	0.24407	0.0310
Corn2-Starch	0.1685	0.0416	0.6669	0.0867	0.6706	0.0818	0.6726	0.0850	0.6776	0.0883
Corn3-Moisture	0.0220	0.0063	0.1070	0.0147	0.1105	0.0147	0.1254	0.0167	0.0997	0.0149
Corn3-Protein	0.0126	0.0037	0.0302	0.0054	0.0257	0.0039	0.0307	0.0051	0.0303	0.0055
Corn3-Fat	0.0354	0.0118	0.2313	0.2957	0.2148	0.0274	0.2344	0.0299	0.2252	0.0284
Corn3-Starch	0.1966	0.0567	0.6980	0.1065	0.6683	0.0987	0.6957	0.1053	0.7023	0.1068

The results of the 15 experiments demonstrate that the DFMA algorithm proposed in this paper outperforms the other four algorithms: In all 15 experiments, the *mean* and *sd* values of the DFMA algorithm were significantly lower compared to its counterpart algorithms. Among them, the results of the fourth set of experiments (corn1-moisture) are the most significant, and the results of the ninth set (corn2-protein) are the closest to the other algorithms. Their corresponding box line plots2&3 are plotted separately as shown below.

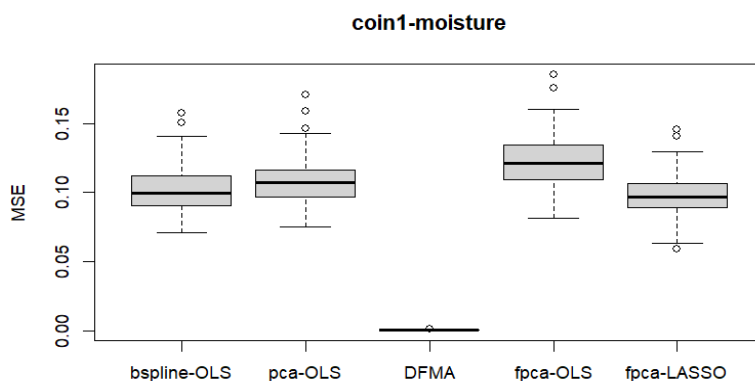


Figure 2 Error box plots of 100 experiments under different comparison methods (corn1-moisture)

The fourth set of experimental data shows that the prediction result of DFMA algorithm is very close to the real value(as shown in Fig.2): according to the 100 sets of experimental data, the *mean* value of DFMA algorithm is 0.0004, and the *sd* value is 0.00018. At the same time, compared with the four comparative methods, DFMA algorithm shows a significant superior performance: in the same set of experiment, the *mean* value of fpca-LASSO is 0.0998, which is the smallest value among the four comparative methods, is 249.5 times of the *mean* value of the DFMA method; the *mean* value of the fpca-OLS method is 0.1262, which is 315.5 times of the *mean* value of the DFMA method. In the ninth group of experiments, the DFMA algorithm still shows a very significant advantage over the other four comparison methods (as shown in Fig.3): in the same group of experiment, the bspline-OLS algorithm has a *mean* value of 0.0299, which is 2.47 times of the *mean* value of the DFMA method and is the smallest among the four comparison methods; the *mean* value of the fpca-OLS method is 0.0306, which is the largest among the four comparison methods, and is the largest among the four comparison methods. The fpca-OLS method *mean* value is 0.0306, the largest among the four comparison methods, which is 2.53 times of the *mean* value of the DFMA algorithm. In conclusion, the DFMA algorithm in the experiments outperforms the four comparative algorithms significantly. This is because the DFMA algorithm has the ability to incorporate the functional data shape information into the regression, and at the same time balance the degree of interpretation of different vectors in the regression after the K-L expansion.

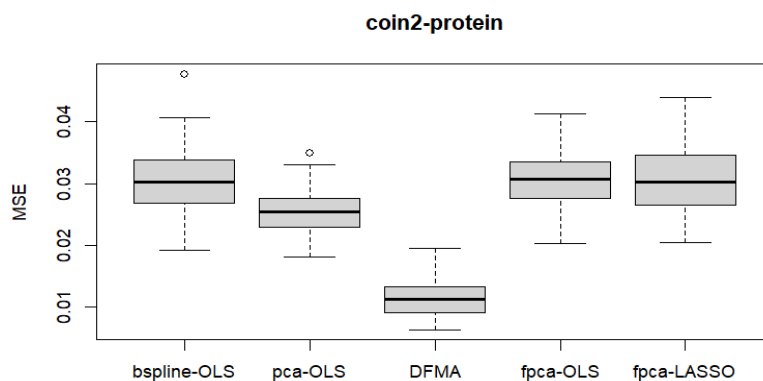


Figure 3 Error box plots of 100 experiments under different comparison methods (corn2-protein)

4. Conclusions

This paper introduces the use of first-order and second-order derivative information of functional data in the regression analysis of functional data on scalars. This approach allows for the shape information of functional data to be fully utilized in order to explain the explanatory variables. It is the first time that such an approach has been applied; meanwhile, Bayesian model averaging better balances the degree of explanation of truncated vectors after KL decomposition on functional data and explanatory variables, and circumvents the problem of under- or overfitting of regression models to a certain extent. This regression method deeply mines and extracts the intrinsic features of the data, making its regression prediction results closer to the real values. In the experiments, this paper conducted 100 experiments on each of the 15 sets of data between the DFMA algorithm and the comparison method, and the algorithm performance was evaluated using the mean and standard deviation of the mean squared error between the true and predicted values across 100 experiments. The experimental results show that this paper's algorithm presents significantly better indicators than the other four algorithms in each of the 15 groups of experiments, reflecting the strong accuracy of the DFMA algorithm.

Regarding future research directions for the use of model averaging methods, it should be noted that the Bayesian model averaging method is limited to linear models. However, there are other model averaging methods with stronger applicability that can be used. For instance, the bagging model does not impose any restrictions on the basic model being a linear model, which significantly enhances the model's generalization ability.

References

- [1] Conway J B. A course in functional analysis[M]. Springer, 2019.
- [2] DING H, XU W C, ZHU H B, et al. Review of regression analysis for functional data[J]. Chinese Journal of Applied Probability Statist, 2018, 34(06): 630-654. (in Chinese)
- [3] Tong H, Ng M. Analysis of regularized least squares for functional linear regression model[J]. Journal of Complexity, 2018, 49: 85-94.
- [4] Likhachev D V. On the optimization of knot allocation for B-spline parameterization of the dielectric function in spectroscopic ellipsometry data analysis[J]. Journal of Applied Physics, 2021, 129(3).
- [5] Hooker, Giles, Hall, et al. Truncated linear models for functional data[J]. Journal of the Royal Statistical Society, Series B. Statistical Methodology, 2016, 78(3): 637-653.
- [6] Zhou H, Yao F, Zhang H. Functional linear regression for discretely observed data: from ideal to reality[J]. Biometrika, 2023, 110(2): 381-393.
- [7] Su, Y.R., Di, C.Z., Hsu, L. Hypothesis testing in functional linear models[J]. Biometrics, 2017, 73, 551-561.
- [8] Zhou, Hang, Fang Yao, and Huiming Zhang. Functional linear regression for discretely observed data: from ideal to reality[J]. Biometrika, 2023: 381-393.
- [9] Aguilera-Morillo M C, Buño I, Lillo R E, et al. Variable selection with P-splines in functional linear regression: Application in graft-versus-host disease[J]. Biometrical Journal, 2020, 62(7): 1670-1686.
- [10] Feng Da-wei. Empirical analysis of cold chain logistics of agricultural products in Shandong province based on principal component analysis and multiple regression[J]. Operations Research and Fuzziology, 2023, 13(5): 4406-4413.
- [11] BAI D F, XU X, WANG G C. Review of generalized linear models and classification for functional data[J]. Journal of Guangxi Normal University (Natural Science Editon), 2022, 40(1): 15-29.
- [12] XIE HE-liang, HU Di. The Application of Multi-factor Quantization Model in Portfolio: The Comparative Research on LASSO and Elastic Net[J]. Journal of Statistics and Information, 2017, 32(10): 36-42.