

Functional Principal Component Regression Model Based on First-order Boosting and Model Averaging

Yifan Huang

School of Computer Science, Northeast Electric Power University, Jilin, 132011

15958594199@163.com

Abstract. Functional regression model has always been an important research direction in the field of functional data analysis, in which functional principal component analysis as a common method of functional regression modeling, the determination of the number of truncations of the eigenfunctions has always been a problem to be solved. Based on this, this paper proposes a functional principal component regression model based on the idea of model averaging, which uses Bayesian model averaging instead of model selection, determines the number of truncations adaptively and balances the prediction error with the training error. In addition, the original function curves do not effectively capture the information of their functional characteristics. In this paper, the first-order derivative function is used as the supplementary information of the original function curve to jointly characterize the predictor variables. In the efficiency test, this paper averages the Bayesian model as the benchmark model and conducts real data analysis on several benchmark datasets. The results show that the proposed method has higher predictive accuracy and robustness than commonly used models.

Keywords: Functional regression, functional principal component analysis, truncation number, Bayesian model averaging.

1. Introduction

With the continuous development of modern data collection and storage technology, data can be acquired and stored in a faster way and in a larger capacity. Under this circumstance, functional data has become a more important and widespread type of data, and the statistical method for studying functional data is generally called functional data analysis model. Specifically, this paper considers the case where the predictor variable is a random function and the response variable is a continuous scalar. This is obviously a typical regression problem. However, when we apply some multivariate regression models directly to function-based data regression, a series of problems such as dimensionality catastrophe, inability to effectively capture function feature information, and easy overfitting of the model will inevitably occur.

When dealing with function-based data regression problems, traditional multivariate analysis methods include principal component analysis, LASSO, tree modeling, and penalized regression. For example, Gao, Xing, Luo, Zhang, & Guo [1] found in their study that PCA is optimal for dimensionality reduction compared to some other algorithms. However, these methods will have the problem of dimensional catastrophe, which will trigger a significant increase in the complexity of computation. Zhang & Fan [2] found that the LASSO regression model shows strong generalization ability and better prediction effect. However, in high-dimensional spaces, especially with nonlinear features, LASSO may have difficulty in correctly capturing the true relationship between the correct response variable and the feature. If the discrete observations are regarded as discrete realizations of random functions, the adaptive penalized spline regression model with local weight adjustment proposed by Zhang & Fan [3], which is highly flexible, interpretable and adaptable, can capture the function feature information while smoothing the data. However, the selection of the number of b-spline bases usually requires some experience and practice, and the selection of different numbers of bases may lead to overfitting or underfitting problems. In addition, besides the b-spline method, such as Fourier basis function, Gaussian basis function, etc. can also be used to smooth the discrete observation points, however, the results obtained by choosing different methods are inconsistent, in

a word, there is no unified paradigm for the selection of the type and number of basis functions. Mu, Wang & Feng [4] used functional principal component analysis (fPCA) to deal with functional data. Specifically, in the face of functional data, the advantages of functional principal component analysis are that it can better deal with functional data, avoid the dimensionality disaster, and can effectively capture the functional features, and the approximation function obtained by functional principal component analysis converges to the real random function according to the probability. The theoretical and computational advantages of functional principal component analysis have led to its use as a common initial step in functional regression models. However, when determining the number of truncated principal components in functional principal component analysis, the usual method is to truncate them based on the cumulative contribution rate. For example, Li, Xie, Chen, Yang, Yang & Jiang [5] used the cumulative contribution rate to select the number of principal components to be retained in the functional additive model. However, when the principal components that have an important influence on the response variable Y appear in the position of lower contribution rate, then if we rely on the cumulative contribution rate to truncate, it will inevitably make the fitted model ineffective.

In order to avoid the appeal problem and further extract the characteristic information of functional data. The purpose of this paper is to propose a functional principal component regression model (DFMA) based on derivative function information and Bayesian model averaging [6] for the regression problem of functional data. The innovation of this method is that on the one hand, the derivative information is used to increase the available domain of information extraction for the original function, and on the other hand, the model averaging method is used instead of model selection, which alleviates the situation that a single prediction model may fall into the overfitting or underfitting situation, and improves the model's generalization ability. Finally, experiments on real data show that the proposed method has ideal application value in the practical problems of functional data regression.

2. Theory and methodology

2.1. Principal component analysis of functions based on derivative information

In order to better capture the function feature information, first-order differencing is first performed on the raw data:

$$X'(t) = \frac{dX(t)}{dt} \quad (1)$$

Let $X(t)$ be the original functional type data, where t denotes the time and $X'(t)$ denotes the data after differencing. And then the B-spline basis function is calculated by Cox-de Boor formula as follows:

$$B_{i,p}(t) = \frac{(t - t_i)B_{i,p-1}(t) + (t_{i+p} - t)B_{i+1,p-1}(t)}{t_{i+p} - t_i} \quad (2)$$

Where: t is a variable on the definition domain, t_i is the i th node in the sequence of nodes, p is the order of the B-spline, $B_{i,p}(t)$ is the i th node and the B-spline basis function of order p , $B_{i,p-1}(t)$ is the i th node and the B-spline basis function of order $p-1$, $B_{i+1,p-1}(t)$ is the $i+1$ th node and the B-spline basis function of order $p-1$. After smoothing the data, the covariance function $Cov[X'(s) - X'(t)]$ is calculated for the processed data as a way to characterize the correlation of two random functions at different points in time, and the covariance function is calculated as follows:

$$Cov[X'(s) - X'(t)] = E[(X'(s) - \mu(s))(X'(t) - \mu(t))] \quad (3)$$

Where: s, t denotes two different time points, $X'(s), X'(t)$ denotes the first order derivative function of the observed data at the two time points, and $\mu(s), \mu(t)$ denotes the mean of the first order derivative function of the observed data at the two time points. The eigenvalue decomposition of the covariance function is performed next:

$$\text{Cov}[X'(s), X'(t)] = \sum_{i=1}^{\infty} \lambda_i \cdot \Phi_i(s) \cdot \Phi_i(t) \quad (4)$$

A set of eigenvalues λ_i and the corresponding eigenfunctions $\Phi_i(t)$ are obtained, the eigenvalues indicate the degree of change in the data, and the eigenfunctions describe the main trends in the data, and then the first-order derivative function $X'(t)$ of the observed data is projected into the space constituted by the eigenfunctions to compute the principal component scores Z_i , which are expressed as:

$$Z_i = \int_a^b X'(t) \cdot \Phi_i(t) dt \quad (5)$$

Where a, b denotes the start time and end time on the definition domain. Then, the feature information of the original data is reconstructed by splicing the principal component sequences of the retained derivative function with those of the original function and saved in the matrix $M[n, h]$, where n denotes the number of samples and h denotes the total number of principal components selected for retention.

2.2. Functional Principal Component Regression Models Based on Bayesian Model Averaging

After calculating the principal component score matrix $M[n, h]$, for each sample, search for each possible linear regression model. At this time, the number of principal components is h , then the candidate model should be $2^h - 1$, for the subsequent formula expression is simple, set the number of candidate models $N = 2^h - 1$. At this time, set the posterior probability of the i th model $\Pr(N_i | D)$ is N_i , where D as the observed data, then the prediction of the average Bayesian model can be expressed as:

$$\hat{Y} = \sum_{i=1}^N \Pr(N_i | D) \cdot \hat{Y}_i \quad (6)$$

$\hat{Y}$ denotes the predicted value of the average Bayesian model, $\hat{Y}_i$ denotes the predicted value of the i th model, and $\Pr(N_i | D)$ denotes the posterior probability of the i th model. And the posterior probability is obtained by Bayes' theorem:

$$\Pr(N_i | D) = \frac{\Pr(D | N_i) \cdot \Pr(N_i)}{\sum_{i=1}^N \Pr(D | N_i) \cdot \Pr(N_i)} \quad (7)$$

Where $\Pr(N_i)$ denotes the prior probability, i.e., the probability of choosing the i th model in the absence of observations; $\Pr(D | N_i)$ denotes the probability of observing data D given model N_i , which can be estimated using the Bayesian information criterion BIC , and BIC can be expressed as:

$$BIC = -2 \cdot \log(L) + k \cdot \log(n) \quad (8)$$

Where L denotes the value of the likelihood function of the model; k denotes the number of parameters of the model; and n denotes the sample size. In addition, the predicted value $\hat{Y}_i$ of the i th model can be expressed as:

$$\hat{y}_i = X\beta_i \quad (9)$$

Where X denotes the input data matrix, i.e., the principal component scores contained in this combination of the independent variables, and β_i denotes the vector of parameter estimates for the i th model. The resulting calculation yields the average Bayesian model prediction, which, when combined with the dependent variable of the observed data, yields the prediction error.

3. Data analysis

3.1. Data sources and experimental setup

In this paper, experiments are conducted on two datasets, the meat dataset and the corn dataset, respectively. First, the meat dataset is available from <http://lib.stat.cmu.edu/datasets/tecator>. This dataset has 240 samples, each containing the moisture, fat, and protein content of the meat, as well as absorption spectra measured with a near-infrared (NIR) spectral analyzer in the range of bands 850-1050 (nm) at intervals of every two bands, with a total of 100 absorption spectrum data recorded. Second, the dataset of corn samples can be downloaded from <http://www.eigenvector.com/data/Corn/index.html>. The dataset was obtained from Cargill USA. The dataset has 80 samples, each containing the moisture, protein, oil, and starch content of the corn, as well as a total of 700 absorption spectrum data measured with a near-infrared spectral analyzer in the band range of 1100-2498 nm at 2-nm intervals. This is shown in Figures 1-2:

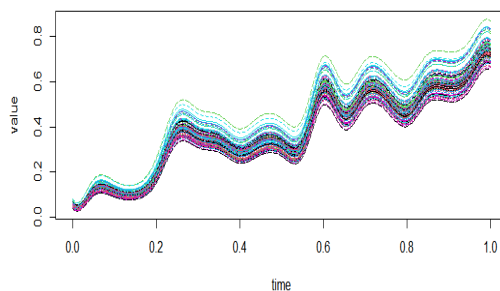


Figure 1. Spectral curves of corn data

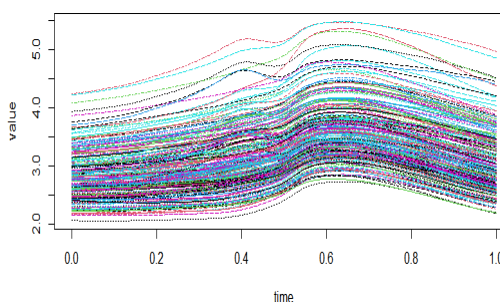


Figure 2. Spectral curves of meat data

The language used for the experiments is R and the tool used is RStudio. For the experiments, in addition to using the method of this paper (DFMA), B-spline and OLS based regression (bspkine-OLS) [7], principal component analysis and OLS based regression (pca-OLS) [8], functional principal component analysis [9] and OLS based regression (fpca-OLS), functional principal component analysis and LASSO based regression (fpca-LASSO) [10]. The experiments were divided into 100 simulations, while the dataset was divided into training and test sets and both were randomly assigned. The ratio of training set to test set was 2:1 when experimenting with meat dataset and 3:1 when using corn dataset. Regarding the calculation of the final mean square error, for each method, the squared error between the predicted value and the true value of the test set was first calculated in each cycle, and the squared error obtained from each cycle was averaged, which means that the mean square error

of each method was calculated. The mean and standard deviation of the mean square error were obtained through 100 cycles. For the results of each experiment another box-and-line plot was created containing the distribution of the prediction errors for the five methods to visualize the comparison results and to assess the stability of the various methods.

3.2. Comparative results

The 100 experimental means and corresponding standard deviations of the mean square error between the predicted and true values of the proposed method and the four comparison methods on the test set are shown in Tables 1-7 below. Specifically, the experimental results for the meat dataset are shown in Tables 1-3. It is evident from Tables 1, 2 and 3 that DFMA has the highest prediction accuracy, i.e., the smallest mean square error, while fpca-OLS and fpca-LASSO perform the worst. As far as the results in Table 1 are concerned, DFMA improves 64.9% over bspline-OLS, 57% over ppca-OLS, 75.7% over fpca-OLS, and 75.7% over fpca-LASSO in terms of mean error.

Table 1. Mean square error of predicted protein for meat dataset

Methods	bspline-OLS	pca-OLS	DFMA	fpca-OLS	fpca-LASSO
mean	2.261918	1.844233	0.793586	3.259699	3.260425
sd	0.525814	0.281504	0.115197	0.795793	0.792665

Table 2. Mean square error of predicted moisture for meat dataset

Methods	bspline-OLS	pca-OLS	DFMA	fpca-OLS	fpca-LASSO
mean	8.020531	8.220872	7.795509	12.798665	12.794118
sd	1.301371	1.466808	1.718436	1.973374	1.961176

Table 3. Mean square error of predicted fats and oils for meat dataset

Methods	bspline-OLS	pca-OLS	DFMA	fpca-OLS	fpca-LASSO
mean	11.352994	11.33108	8.651872	22.061457	22.055267
sd	2.032394	2.09308	1.863839	3.240661	3.235452

Table 4. Mean square error of predicted protein for corn dataset

Methods	bspline-OLS	pca-OLS	DFMA	fpca-OLS	fpca-LASSO
mean	0.12705739	0.08317012	0.020986950	0.21028944	0.21061644
sd	0.03553919	0.02883307	0.009906235	0.04487737	0.04437191

Table 5. Mean square error of predicted moisture for corn dataset

Methods	bspline-OLS	pca-OLS	DFMA	fpca-OLS	fpca-LASSO
mean	0.04946903	0.04619579	0.02282274	0.04990321	0.04938349
sd	0.1418063	0.02134514	0.01081566	0.01369416	0.01369888

Table 6. Mean square error of predicted fats and oils for the corn dataset

Methods	bspline-OLS	pca-OLS	DFMA	fpca-OLS	fpca-LASSO
mean	0.022461418	0.016103210	0.012919262	0.023012063	0.023519861
sd	0.006402905	0.005888546	0.005357874	0.005891858	0.006058575

Table 7. Mean square error of predicted starch for corn data set

Methods	bspline-OLS	pca-OLS	DFMA	fpca-OLS	fpca-LASSO
mean	0.4066028	0.4125663	0.20414915	0.6817283	0.6788256
sd	0.1247307	0.1515649	0.07928576	0.1657703	0.1641186

Similarly, the experimental results for the maize dataset are evident from Tables 4, 5, 6 and 7. the DFMA prediction accuracy remains higher than the remaining four methods. For example, as analyzed from Table 4, DFMA improves 83.5% over bspline-OLS, 74.8% over pca-OLS, 90.0% over fpca-OLS, and 90.0% over fpca-LASSO in terms of error mean.

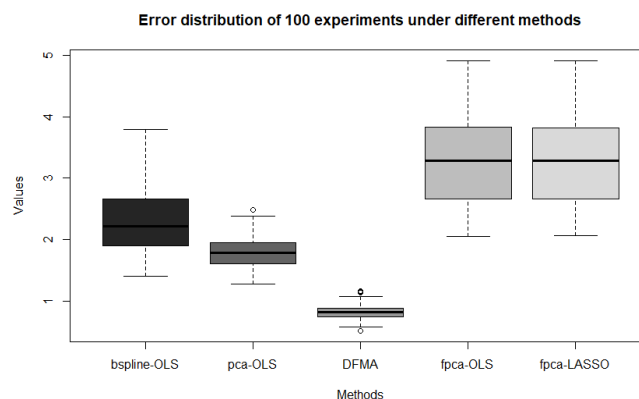


Figure 3. Boxplot of prediction errors regarding proteins in the meat dataset

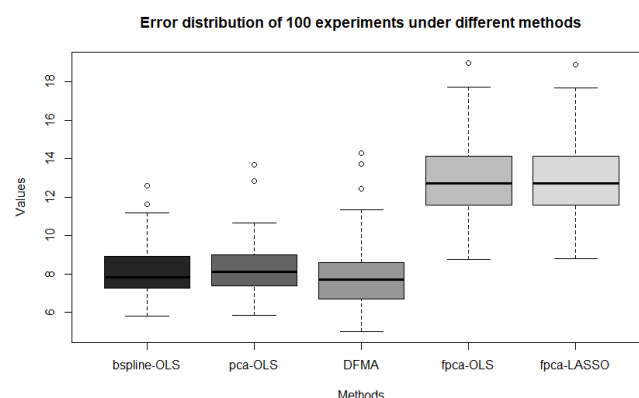


Figure 4. Boxplot of prediction errors regarding moisture in the meat dataset

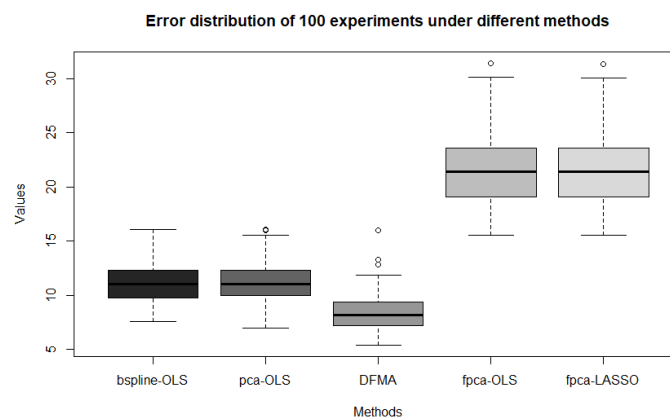


Figure 5. Box plots of prediction errors for fats and oils in the meat dataset

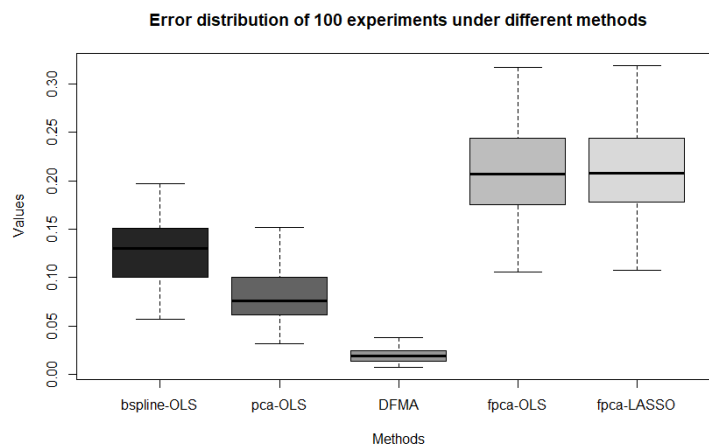


Figure 6. Boxplot of prediction errors regarding protein in the corn dataset

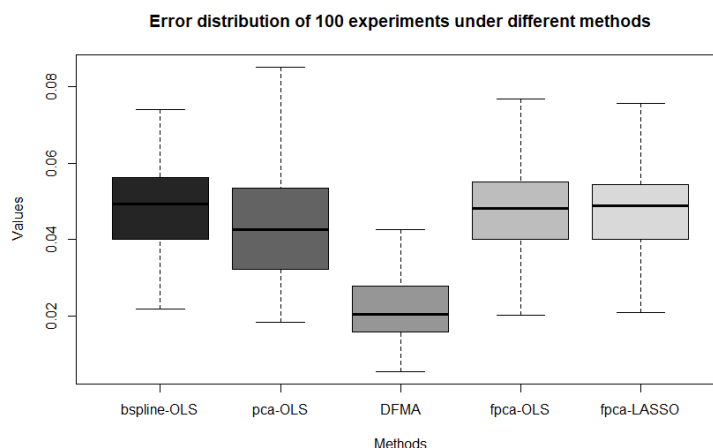


Figure 7. Boxplot of prediction errors regarding moisture in the corn dataset

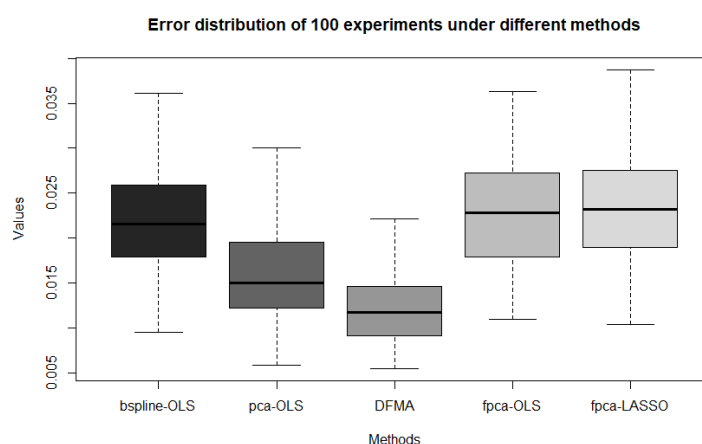


Figure 8. Boxplot of prediction errors on fats and oils in the corn dataset

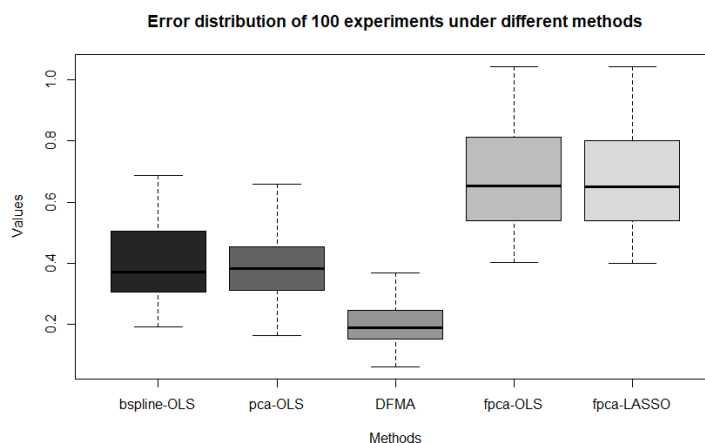


Figure 9. Boxplot of prediction error regarding starch in the corn dataset

Above are the box plots obtained based on the final predictions of the two datasets, and based on these box plots, it is intuitively clear that of the five methods, DFMA is the best in terms of both the magnitude of the prediction error and its robustness

In summary, the method proposed in this paper performs better in prediction accuracy, which is partly due to the fact that we have extracted additional information about the first-order derivative function of the original data, i.e., we have captured more information about the function's features compared to the original data. The other reason is because after obtaining the spliced principal component sequences, we use the performed Bayesian model averaging instead of the truncated model based on cumulative contributions. Compared with a single model, BMA can consider multiple models simultaneously and synthesize the advantages and disadvantages of multiple models by

weighted averaging, which reduces the uncertainty in the model selection process, reduces the risk of overfitting and underfitting, and improves the stability and generalization ability of the model.

4. Conclusion and replication

The first-order derivative information-based principal component regression model (DFMA) based on Bayesian model averaging proposed in this paper achieves high prediction accuracy and robustness in function-based data regression problems. The DFMA method is able to better capture the function feature information by utilizing the first-order derivative function of the original data to perform function-based principal component analysis, which improves the prediction accuracy. The introduction of Bayesian model averaging makes the model more robust and reduces the risk of overfitting and underfitting. The DFMA method performs well in dealing with the regression problem of function-type data, and it can be generalized and applied to other fields such as biomedical and engineering sciences, etc.

The first-order derivative function used in the DFMA as a means of feature extraction can also be applied to other functional data analysis methods, such as clustering, classification and other problems. This feature extraction method can improve the effectiveness of data analysis and the interpretability of the results, so it is of some relevance. By studying its application and effect in different fields, it can further expand its application scope in practical problems and improve the accuracy and stability of data analysis. In summary, the successful application of the DFMA method in function-based data regression problems not only improves the prediction accuracy and robustness, but also provides new ideas and methods for function-based data analysis methods in other fields.

References

- [1] Gao, F., Xing, Y., Luo, H., Zhang, Y., & Guo, L. (2024). Non-destructive discrimination method of apricot varieties based on visible/near-infrared spectroscopy and chemometrics. *Spectroscopy and Spectral Analysis*, 44 (1), 44-51. Zhang, H., & Fan, M. (2022). Analysis of Qingdao's marine economy and marine industry based on the Lasso regression model. *Ocean Development and Management*, 39 (08), 22-28.
- [2] Yin, Y., Yang, L., & Wang, X. (2018). Adaptive penalized spline regression model with local weighting adjustment. *Applied Mathematics*, 31 (3), 592-599.
- [3] Mu, W. Y., Wang, X. Y., & Feng, Z. H. (2023). Functional data outlier detection for spectral data. *Chinese Journal of Applied Probability and Statistics*, 39 (5), 667–681.
- [4] Li, M., Xie, X., Chen, Z., Yang, M., Yang, D., & Jiang, J. (2018). Monitoring of Semiconductor Manufacturing Process Based on Functional Data Analysis. *Journal of Mechanical Engineering*, 54 (16), 62-69.
- [5] Fragoso, T. M., Bertoli, W., and Louzada, F. (2018) Bayesian Model Averaging: A Systematic Review and Conceptual Classification. *International Statistical Review*, 86: 1–28.
- [6] Liu, J., Yi, S., Yao, J., et al. (2017). Research on video defogging based on least squares B-spline curve. *Journal of Changshu Institute of Technology*, 4, 50-55.
- [7] Yuan, S., Zhang, J., Liu, M., Zhang, J. (2016). Research on cherry sweetness detection based on principal component regression analysis and partial least squares regression. *Journal of Analytical Science*, 32 (4), 553-556.
- [8] Su, B., Jiang, J., Tang, Q., & Sheng, M. (2017). Human dynamic behavior recognition based on functional data analysis methods. *Acta Automatica Sinica*, 43 (5), 866-876.
- [9] Huang, Q., Zheng, S., Zhang, Z., Zhu, D., Fan, X., Du, B., & Liu, S. (2024). Construction of a risk prediction model for metabolic syndrome based on Lasso regression using biomarkers. *Chinese Journal of Rehabilitation Medicine*, 33 (1), 1-5.
- [10] Zhang, X. Y., & Zou, G. H. (2011) Model Averaging Method and Its Application in Forecast. *Statistical Research*, 28, 97-102.

- [11] Zhang, X. Y. Model averaging by least squares approximation (in Chinese). *Sci Sin Math*, 2021, 51: 535–548.