Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Functional data regression model based on basis function expansion and Group Lasso

Xuyi Shi *, Jiachen Guang

School of Mathematics and Statistics, Beijing Technology and Business University, Beijing, China, 100048

* Corresponding author: uccshixuyi@163.com

Abstract. In the analysis of functional data, functional data regression is a crucial statistical model. The preceding step is to smooth the discrete function, usually using the basis function expansion method. However, due to the diversity of basic functions, it is a difficult problem to choose the basis function suitable for specific data. In view of this, an adaptive basis function expansion function regression model based on Group LASSO is proposed in this paper, which is suitable for the scenario where the predictor is a random function and the response variable is a continuous scalar. The proposed method uses a variety of basis function expansion methods, significantly expands the function space, and can smooth complex random functions better than a single basis function expansion method. On the other hand, to avoid the overfitting problem caused by the over-complexity of the prediction function, this paper adopts the Group LASSO method to automatically select the best type of basis function to alleviate the overfitting problem. The effectiveness of the proposed method is verified by an empirical study on the NIR reflectance spectral datasets of meat and corn samples. The results show that the proposed method has smaller mean square error and absolute error and is robust compared with the functional regression model with single basis function expansion.

Keywords: Functional regression model; Basis function expansion; Group LASSO.

1. Introduction

The concept of functional data was initially proposed by Canadian statistician Ramsay in the 1970s. This concept integrates functional analysis, topology, and statistics, forming the concept of "functional data" and the corresponding data processing method called Functional Data Analysis (FDA) [1]. FDA has been successfully applied in various fields, such as the analysis of bone shape in archaeology, recording time series in economics, tracking the movement trajectory of handwriting pen tips, temperature changes, and changes in human height [2]. In the context of functional data processing, functional data regression models have become an important class of models. Taking the dataset of meat and corn absorbance spectra as an example, this model utilizes the absorbance spectra curves of meat and corn as predictor variables to respectively predict the water, fat, and protein content of meat, as well as the oil content, moisture content, starch content, and protein content of corn.

In the analysis of such data sets, multiple linear regression models are often used to fit each content in the sample. However, due to the large number of discrete observation points in the spectral curve, the dimension of the corn data set, for example, is as high as 700 dimensions, which can lead to dimensional disasters when applying multiple linear regression models. There are many solutions to the dimensional disaster problem, such as LASSO regression and ridge regression, dimensionality reduction by adding regularization, or random forest methods. However, these methods often treat each dimension as an independent data point and fail to take into account the correlation between the dimensions, thus ignoring the characteristic information of the function itself. Therefore, functional regression models are more suitable for dealing with such data sets. In general, basis function expansion is chosen to solve such models. For example, Mikael Manngard and Hannu T. Toivonen used Laguerre basis function expansion to identify low-order models [3]. Jerry M. Mendel proposed

the application of adaptive variable structure basis function expansion in machine learning [4]. Tomohisa O, Yukitoshi F and Takuya N used ABIC basis function expansion to estimate the strain rate field of GNSS data [5]. Li and Hsing used Fourier basis functions to approximate functional variables and function coefficients, and proposed a least-squares estimation with penalty [6]. Zhou and Wang, among others, explored the relationship between variables by using B-spline expansion for functional data sets [7]. However, one issue that needs to be solved is how to choose a basis function that is more suitable for a particular situation.

Based on this, considering that the predictor is a random function and the response variable is a continuous scalar, a functional regression model using Group LASSO method is proposed to achieve adaptive basis function expansion. Among them, Group LASSO is a regularization technique for feature selection and model parameter estimation. Compared with general LASSO regression, Group LASSO allows groups of related covariate sets to be selected as a single unit, which can be useful in settings where it does not make sense to include some covariate sets without others. [8]. The method in this paper assigns different weight coefficients to different basis functions and combines them into a new linear combination. By making the correlation coefficient randomly zero, the purpose of selecting the appropriate basis function is achieved. As a result, Group LASSO is able to pick out the most suitable basis function or multiple basis functions. Empirical data analysis shows that the proposed method has smaller mean square error and absolute error and is more robust than the single basis function expansion.

2. Functional data regression model based on basis function expansion and Group LASSO

Generally speaking, in a linear regression model, a continuous response variable Y can be expressed as

$$Y = \int x(t)\beta(t)dt + \varepsilon \tag{1}$$

Where t represents the value observed at a certain time, $x(t)$ represents the value observed at time t, $\beta(t)$ represents the weight of the observed value at time t, and ε represents the residual. Suppose we first get the J - dimensional vector $x(t)$ in the form of discrete points, then we smooth the discrete points so that $x(t)$ can be expressed as a basis function expansion

$$x(t) = \sum_{j=1}^{J} \alpha_j \varphi_j(t) \tag{2}$$

Where α_j is the linear combination coefficient, which can be obtained according to the principle of least squares

$$\alpha_j = (x^T x)^{-1} x^T y \tag{3}$$

Here x is the eigenvector of $\varphi_j(t)$, and y is the value corresponding to the discrete point of $x(t)$. So $\varphi_j(t)$ is the basis function. There are many types of basis functions. This paper takes functional principal component analysis basis function, Fourier basis function and B-spline basis function as examples to select the basis function.

(1) Functional principal component basis function

The basis function is expanded according to Karhunen-Loeve [9]

$$x_i(t) = \mu_i(t) + \sum_{i=1}^{\infty} \alpha_{ik}\varphi_k(t) \tag{4}$$

Where

$$\alpha_{ik} = \int_I (x_i(s) - \mu(s))\varphi_k(s)ds \tag{5}$$

(2) Fourier basis function

According to the Fourier series with the period T as an example, we obtain

$$\varphi(t) = \frac{\alpha_0}{2} + \sum_{k=1}^{n}[\alpha_k \cos(\frac{2k\pi}{T}t) + \beta_k \sin(\frac{2k\pi}{T}t)] \tag{6}$$

Where

$$\alpha_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \varphi(t)cos(kt)dt \tag{7}$$

$$\beta_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \varphi(t)sin(kt)dt \tag{8}$$

(3) Spline basis function

Let $U = \{u_0, u_1, ..., u_m\}$ be a monotonically undecreasing sequence of real numbers, which is $u_i \le u_{i+1}$, $i = 0,1,2,...,m-1$. Where u_i is called the node, U is called the node vector, and $N_{i,p}(u)$ is used to represent the ith p degree ($p+1$ order) B-spline basis function

$$\begin{aligned} N_{i,p}(u) &= \begin{cases} 1, & u_i \le u < u_{i+1} \\ 0, & else \end{cases} \\ N_{i,p}(u) &= \frac{u - u_i}{u_{i+p} - u_i} N_{i,p-1}(u) + \frac{u_{i+p+1} - u}{u_{i+p+1} - u_{i+1}} N_{i+1,p-1}(u) \end{aligned} \tag{9}$$

Next, combine Group LASSO with the basis function expansion. Specifically, in a linear regression model, in the case of formula (1), the Group LASSO estimate can be derived as

$$\hat{\beta}_\lambda = arg\min(\|Y - X\beta\|_2^2 + \lambda\sum_{g=1}^{G}\left\|\beta_{I_g}\right\|_2) \tag{10}$$

Where I_g refers to the subscript set of group g variables, and β_{I_g} refers to the coefficient variable of group g variables. Because the penalty term of Group LASSO method can be regarded as the middle state of L1 penalty and L2 penalty [10]. This method selects variables at the group level,

that is, the selection variables of groups [11], so we can group all variables, that is, $\beta = \{\beta_1, \beta_2, ..., \beta_n\}$, and then punish the L2 norm of each group in the objective function, so that a whole set of coefficients can be eliminated to 0 at the same time. This means that we need to group the four basis functions, the B-spline basis function named Φ_1, the FPCA(Functional Principal component analysis) basis function named Φ_2, the Fourier basis function named Φ_3, and combine them into a new coefficient vector

$$\beta = \begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \Phi_3 \end{pmatrix} \tag{11}$$

Finally, by using the selection method of Group LASSO, Φ_1 is taken as Group 1, Φ_2 as Group 2, Φ_3 as Group 3, and vector β is brought into the model for solving.

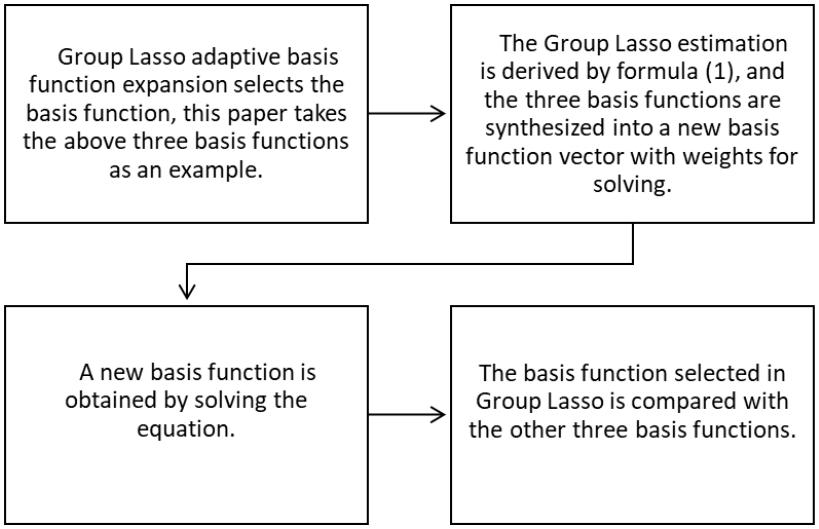

Figure 1. Steps for Group LASSO to select a basis function

3. Data analysis

3.1. Data sources and Experimental Settings

This paper verifies the model effect by testing on two classical functional datasets. First, we use the meat sample spectra data set, the Numbers from http://lib.atat.cmu.edu/datasets/tecator. This dataset has also been cited in the reference for other functional data, such as Ferraty and Vieu [12], Aneiros-Pérez and Vieu [13].The dataset has 240 sample data, each recording discrete absorption spectral curves in the 850-1050 nanometers (nm) range of the near infrared spectrum analyzer, with 100 different absorption spectral sample points within each of the two intervals, as well as the fat, protein, and water content of each sample. Our prediction task is to use 100 samples of meat samples as input information $x_i(t)$ to predict the fat content, protein content and water content of meat respectively and record it as Y, where the value of t is in the interval of 850-1050 nm. Secondly, we use the samples of corn spectral data set, and the data can be downloaded from http://www.eigenvector.com/data/Corn/index.html. This data set is composed of 80 samples, each of which records the interval between 1100-2498 nm. Each corn sample had 700 different absorption spectrum sample points at interval of every 2 nm . Our same prediction task is to predict the water content, oil content and protein content of corn respectively based on the sample data of corn and record it as Y, while 700 samples are used as input information $x_i(t)$, and the value of t is in the

interval of 1100-2498 nm. These data are discrete functional data, which usually requires a pre-processing step to transform them into continuous curves, and our proposed method is used to deal with this situation. In the comparison, we will use the method proposed in this paper to compare with some common pretreatment methods (smooth method) to evaluate its advantages and disadvantages. Such a comparison will contribute to a better understanding of the effectiveness and practicability of the proposed method when dealing with functional data.

Programming in R was used as experimental software in the experiment. In order to compare the performance of the model used in this paper compared with B-spline basis function model, FPCA basis function model and Fourier basis function model, we randomly divided the data into training set D_1 and test set D_2. In D_1, 155 samples were randomly selected, and the remaining samples were taken as data in test set D_2. On the test set, mean square error and absolute error analysis were carried out:

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{13}$$

And then, the experiment was repeated 100 times for getting the mean and standard deviation of 100 results.

3.2. Comparison result

On the one hand, after predicting the fat, protein and water content of meat respectively, we summarized the results obtained by the four methods to get a box chart 2-4 and Table 1 of mean square error and absolute error under 100 experiments. On the other hand, after predicting the water content, oil content and protein content of corn respectively, we summarized the results obtained by the four methods, and got a box chart 5-7 and Table 2 of mean square error and absolute error under 100 experiments.

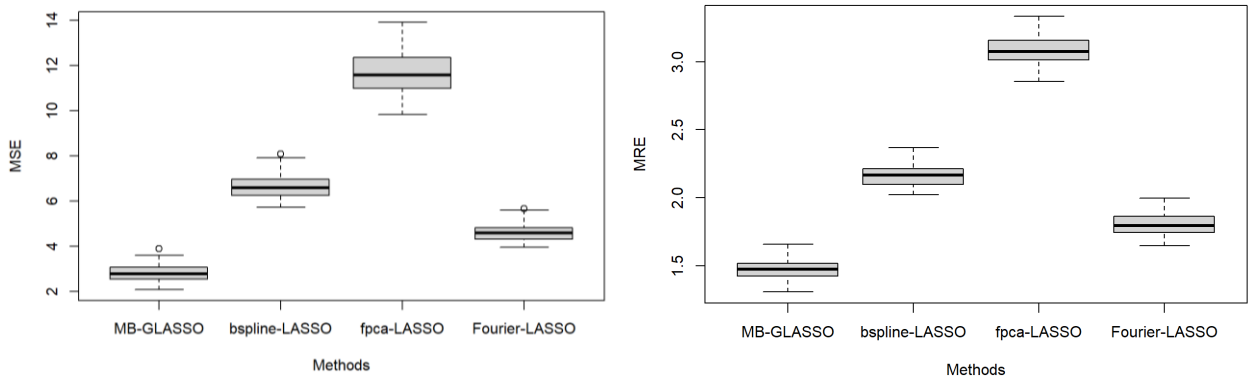

Figure 2. Comparison of MSE and MRE for predicting Meat Fat content

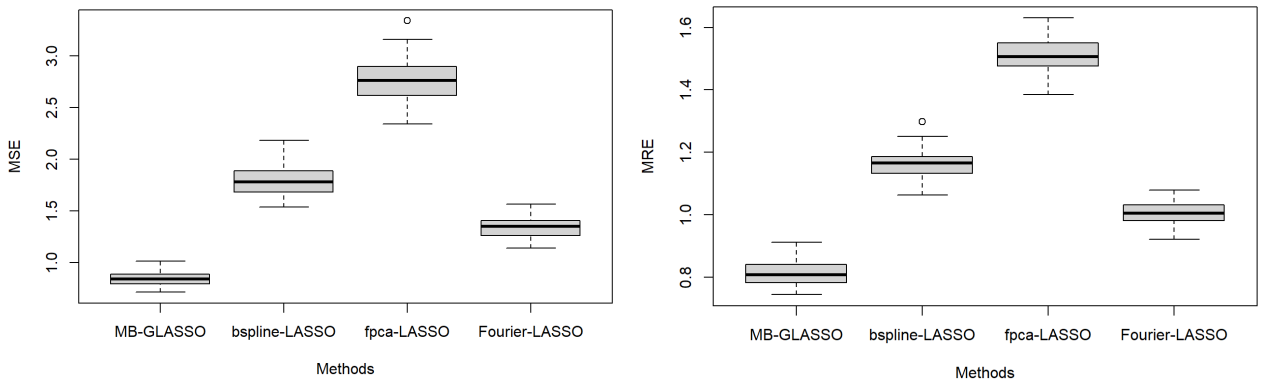

Figure 3. Comparison of MSE and MRE for predicting Meat Protein content

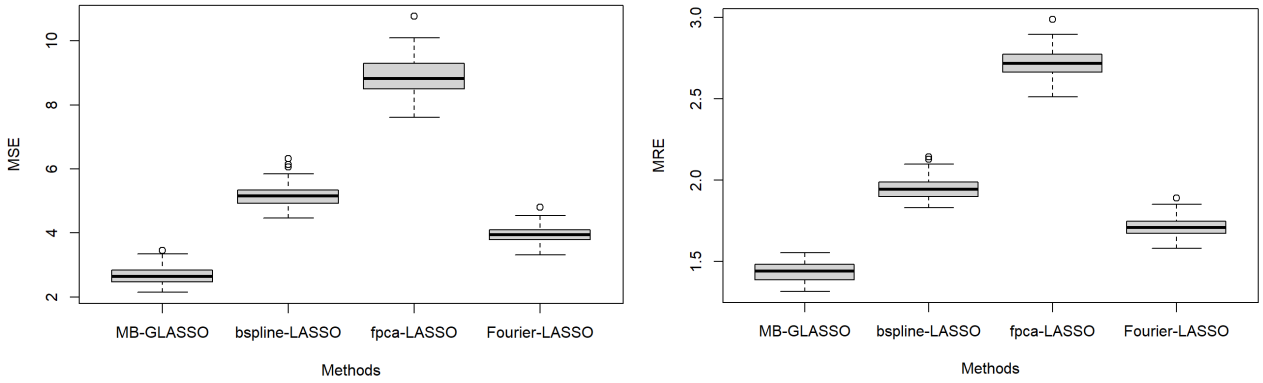

Figure 4. Comparison of MSE and MRE for predicting Meat Water content

Table 1. Mean square error and absolute error of meat data

Meat	Method		MB-GLASSO	bspline-LASSO	fpca-LASSO	Fourier-LASSO
Fat	MSE	mean	2.8171923	6.629183	11.6955404	4.5985850
		sd	0.3715523	0.504011	0.8913739	0.3724526
	MRE	mean	1.47160605	2.16866819	3.0867114	1.80299939
		sd	0.06986228	0.08272987	0.1047226	0.07550786
Protein	MSE	mean	0.84659201	1.7859512	2.7596945	1.34139275
		sd	0.07211732	0.1344351	0.1844491	0.09955042
	MRE	mean	0.81221320	1.15950630	1.51132225	1.00502876
		sd	0.03488036	0.04253976	0.05347183	0.03422928
Water	MSE	mean	2.6747842	5.1572870	8.8771862	3.9538837
		sd	0.2719448	0.3476811	0.6140655	0.2436538
	MRE	mean	1.43669634	1.94662616	2.71644291	1.71616387
		sd	0.05966421	0.06437473	0.08812882	0.05661001

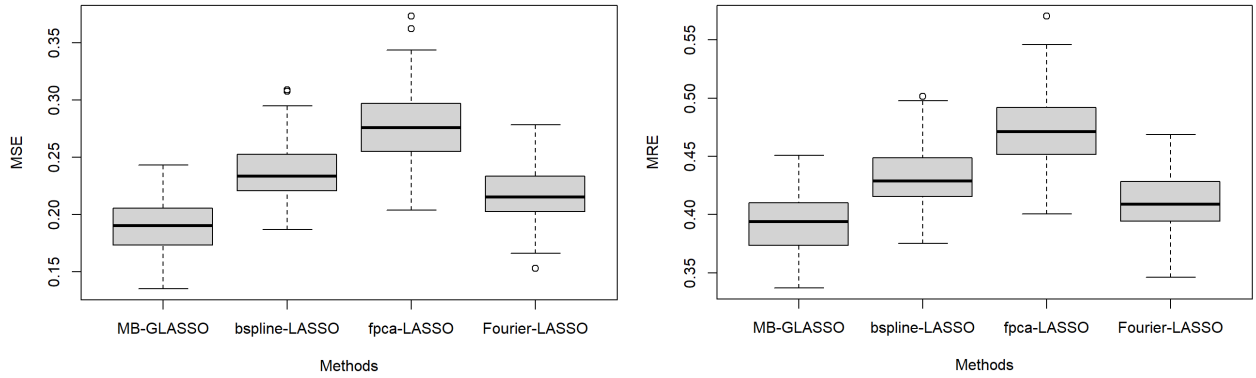

Figure 5. Comparison of MSE and MRE for predicting Corn Moisture content

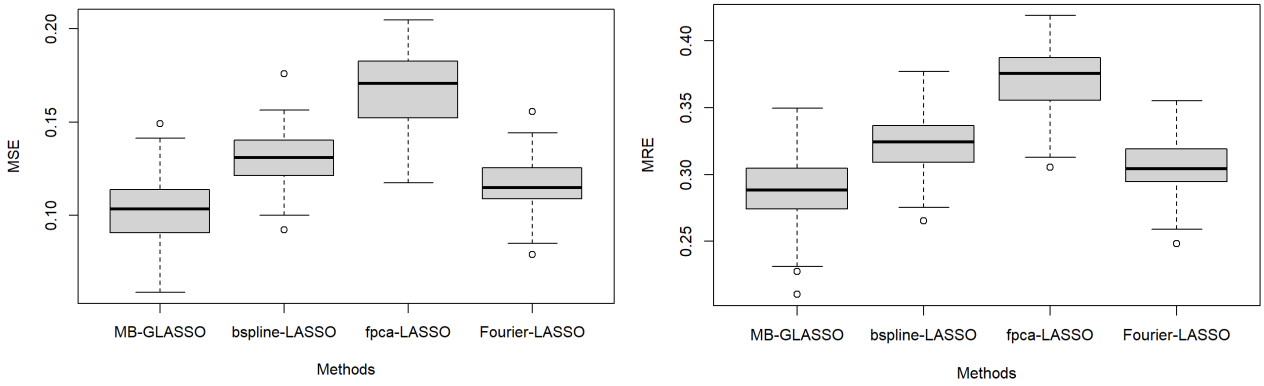

Figure 6. Comparison of MSE and MRE for predicting Corn Oil content

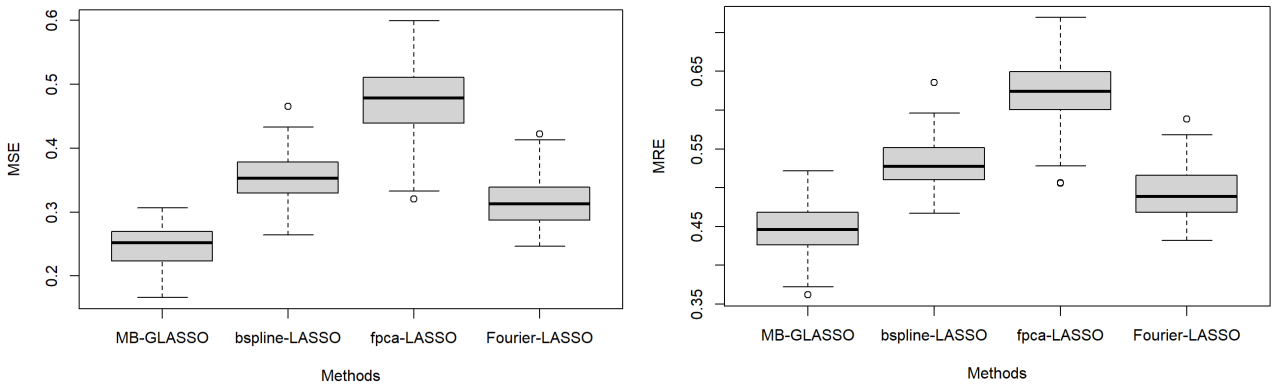

Figure 7. Comparison of MSE and MRE for predicting Corn Protein content

Table 2. Mean square error and absolute error of corn data

Corn	Method		MB-GLASSO	bspline-LASSO	fpca-LASSO	Fourier-LASSO
Moisture	MSE	mean	0.18975664	0.23719407	0.27653592	0.21606191
		sd	0.02427562	0.02621142	0.03131186	0.02327073
	MRE	mean	0.39248005	0.43172145	0.47271301	0.40913274
		sd	0.02606836	0.02597107	0.03054662	0.02521088
Oil	MSE	mean	0.10249450	0.1311315	0.16763071	0.11658412
		sd	0.01658674	0.0138679	0.02007977	0.01218198
	MRE	mean	0.28614744	0.32217458	0.37141769	0.3055936
		sd	0.02459936	0.01895176	0.02353584	0.0181132
Protein	MSE	mean	0.24634907	0.35481672	0.47550718	0.31540846
		sd	0.03179143	0.03558747	0.05521406	0.03908825
	MRE	mean	0.44698024	0.53044173	0.62353741	0.49425957
		sd	0.03189147	0.02948686	0.04262736	0.03233413

Typically, it can be seen from Figure 2 and Table 1 that the mean and standard deviation (this is represented by “sd” in the table, as is the standard deviation below) of absolute error (MRE) of Group LASSO 's method in predicting meat fat content are relatively small among the four methods. Compared with B-spline basis function expansion, the mean value and standard deviation of this method are lower by 0.697 and 0.013 respectively. Compared with FPCA basis function expansion, the mean value and standard deviation of this method are about 1.615 lower and 0.035 lower. Compared with Fourier basis function expansion, the mean value of this method is about 0.331 lower, but the standard deviation is about 0.006 lower. For another example, it can be seen from Figure 7 and Table 2 that the mean square error (MSE) and standard deviation of the method applied by Group LASSO in predicting corn protein content are relatively small among the four methods. Compared with B-spline basis function expansion, the mean value and standard deviation of this method are lower by 0.108 and 0.004 respectively. Compared with FPCA basis function expansion, the mean value and standard deviation of this method are about 0.229 lower and 0.023 lower. Compared with Fourier basis function expansion, the mean value of this method is about 0.069 lower, but the standard deviation is about 0.007 lower.

Based on these results, among the four different basis function expansion methods, the mean and variance of the mean square error (MSE) and mean relative error (MRE) are relatively low. Firstly, for the MSE, a lower mean of the MSE indicates that the fitted values after 100 fits are closer to the observed values, implying a smaller gap between the fitted data and the real data. Moreover, the standard deviation is also the smallest among the four, indicating that the prediction data based on Group LASSO adaptive basis function expansion is relatively more stable compared to other methods. Similarly, for the mean relative error, the conclusion holds. Therefore, it can be concluded that after selecting the basis functions with Group LASSO, compared to the function-type data regression model based on single basis function expansion, the function space is significantly broadened. This allows for smoothing of more complex random functions, which explains why the mean square error and mean relative error of the function-type data regression model based on Group LASSO adaptive basis function expansion are superior to those of single basis function models. Additionally, it reduces the complexity of the model to prevent overfitting issues, thereby achieving a relatively balanced state in terms of model accuracy and generalization.

4. Conclusion and prospect

This paper proposes a function-type data regression model based on Group LASSO adaptive basis function expansion for the case where the predictor variable is a random function and the response variable is a continuous scalar. Firstly, the model adaptively selects optimal basis functions or combinations of basis functions using Group LASSO, and then expands the function to minimize the

mean square error and mean absolute error between the fitted values and observed values, compared to single basis function expansion. This implies that, on one hand, the model effectively expands the function space through various basis function expansion methods, allowing the model to smooth more complex random functions and better adapt to a wider range of function-type data. On the other hand, the method adaptively determines the optimal types of basis functions through Group LASSO, reducing overfitting issues in the prediction model and improving its accuracy, even surpassing the accuracy of models based on partial single basis function expansion. Additionally, this method ensures that the model has smaller variance and is more stable compared to models based on single basis function expansion.

As for future research directions, besides the Group LASSO method, there are other ways to select models, such as Bayesian model averaging in the model averaging approach. In this experiment, Bayesian model averaging allows multiple models to be combined with weights based on the reciprocals of their respective model residuals, without readily excluding any model, thus effectively addressing the problem of low model fitting caused by missing effective information. With the rapid development of computer technology, model averaging methods, as one of the more complex processing methods, will be more widely used in practical problems.

References

[1] Ramsay J O, Silverman B W. Functional Data Analysis [M]. New York: Springer-Verlag, 2005.

[2] Su Benyue, Chen Xiaohui, Tong Xinghui, et al. Two types of functional data principal component analysis methods and their applications[J]. Statistics and Decision Making, 2015 (17): 24 - 28.

[3] Manngård M, Toivonen T H. Identification of low-order models using Laguerre basis function expansions ∗∗ M.M. gratefully acknowledge the financial support from the Finnish Graduate School in Chemical Engineering (GSCE). [J]. IFAC Papers Online, 2018, 51 (15): 72 - 77.

[4] Mendel M J. Adaptive variable-structure basis function expansions: Candidates for machine learning [J]. Information Sciences, 2019, 496124 - 149.

[5] Tomohisa O, Yukitoshi F, Takuya N. Consistent estimation of strain-rate fields from GNSS velocity data using basis function expansion with ABIC [J]. Earth, Planets and Space, 2021, 73 (1).

[6] Li Y, Hsing T. On rates of convergence in functional linear regression [J]. Journal of Multivariate Analysis, 2007, 98 (9): 1782 - 1804.

[7] Zhou J H, Wang N Y, Wang Functional linear model with zero-value coefficient function at sub-regions [J]. Statistica Sinica, 2013, 23 (1): 25 - 50.

[8] Yuan M, Lin Y. Model Selection and Estimation in Regression with Grouped Variables [J]. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2006, 68 (1): 49 - 67.

[9] BETZ W, PAPAIOANNOU I, STRAUB D. Numerical methods for the discretization of random fields by means of the Karhunen-Loève expansion [J]. Computer Methods in Applied Mechanicsand Engineering, 2014, 271: 109 - 129.

[10] Zou H, Hastie T, Tibshirani R. On the "degrees of freedom" of the LASSO [J]. The Annals of Statistics, 2007, 35 (5): 2173 - 2192.

[11] Wang Hongjie, Wang Ke, Yu Shuixiang, et al. Analysis of key factors in quality formation of Gastrodia elata based on Group-LASSO [J]. Chinese Herbal Medicines, 2023, 54 (13): 4278 - 4285.

[12] Ferraty F, Vieu P. The Functional Nonparametric Model and Application to Spectrometric Data [J]. Computational Statistics, 2002, 17 (4). 545 - 564.

[13] Aneiros-Pérez G, Vieu P. Semi-functional partial linear regression. Statistics & Probability Letters, 2006, 76 (11), 1102 - 1110.