Functional data regression model based on functional mean variance estimation and mutual information weighted ensemble learning

Jianxun Liu[1, *, #], Zijun Liu[1, #], Jiajian Chen[2, #]

[1]School of Minzu University, Beijing, China, 100081

[2]School of Harbin Commercial University, Heilongjiang Province, China, 150028

[*]Corresponding author: ljx620511@163.com

[#]These authors contributed equally.

Abstract. Functional data regression model holds significant application value in various fields such as medicine, economics, finance, and industrial manufacturing. This paper introduces a novel functional regression method that incorporates functional mean variance estimation and mutual information weighted ensemble learning. The proposed method employs functional mean variance estimation to reduce dimension while maintaining the effective information of the predictor related to the response variable. Specifically, this method projects the random function into the function space spanned by finite dimensional basis functions to extract the underlying functional information of the original data. Additionally, considering that the number of sufficient dimension reduction sub-directions is unknown, this paper proposes a model averaging approach based on mutual information weighting to determine the importance of each sub-model. The proposed method can adaptively address the over-fitting and under-fitting problems of the prediction model. The empirical data analysis indicates that the proposed method outperforms other comparison methods in terms of smaller mean square error and absolute error, and it exhibits a certain level of robustness.

Keywords: Functional data regression, Basis function expansion and sufficient dimension reduction, model averaging.

1. Introduction

In recent years, the development of computer and network technology has resulted in an increasing amount of data being presented in the form of curves. These curves include real-time stock prices, daily average temperatures, spectral waves, and ECG curves. This type of data is commonly known as functional data and has important applications in fields such as meteorology, biology, economics, finance, and more. In finance, the efficient processing of high-frequency continuous process data is a crucial issue in financial quantitative research. Similarly, in image recognition, real-time motion trajectory of handwritten numbers can be used to predict specific numbers. In the medical field, brain wave diagrams obtained through brain wave imaging are used to identify brain mechanism patterns. Lastly, in the field of chemistry, the light waves generated by infrared spectrometers are used to measure the chemical substances present in a sample[1-2].

Firstly, functional data can be regarded as high-dimensional data. There are two common methods for processing high-dimensional data, one is generalized linear model, the other is regression based on dimension reduction. In the generalized linear model, the Lasso regression proposed by Robert (1966)[3]is often used to constrain the model by introducing a penalty term. However, the Lasso regression model is easy to cause over-fitting of the model, which makes the prediction model unable to generalize and solve a wide range of problems[4][5]. In the method of regression after dimension reduction, the principal component analysis method proposed by Pearson (1901)[6] is widely used in many scientific fields. However, the number of principal components extracted by principal component analysis is determined by the cumulative contribution rate, and its selection is independent of the response variable y, which cannot guarantee the importance of each component to y. Therefore, using this method will lose the prediction information of the response variable[7]. In contrast, a supervised dimension reduction method called sufficient dimension reduction is more prominent[8]. The core idea of this method is that the sufficient dimension reduction method uses the conditional independence principle without losing information, which can effectively retain the effective information of the original data for the response variable. However it may lose the relationship between the original data if we just regard functional data as high-dimensional data, we should take the functional structure information [9]. In the functional data analysis method, the general method of processing such data is to use the basis function expansion, such as B-spline basis, wavelet basis, Fourier basis, functional principal component basis[10] and so on. However, in the functional data regression problem, using a fixed basis function to expand, although the function information can be extracted, it does not guarantee that the extracted function structure information is related to the response variable. In contrast, the functional sufficient dimension reduction method expands the sufficient dimension reduction idea in the multivariate case[11], which can effectively retain the effective information between the functional random variables and the response variables while reducing the dimension.

In summary, this paper proposes a novel functional data regression model that utilizes functional mean variance estimation and mutual information weighted ensemble learning. The innovation of this method lies in its two key aspects. Firstly, it smooths and projects the observed discrete random function variables into a function space using a reduced-dimensional basis function generated by functional mean variance estimation. This ensures that information in the prediction variable that is valid for the response variable is not lost during dimension reduction. Secondly, it employs the idea of model averaging[12] to adaptively determine the number of functional sufficient dimension reduction sub-directions and effectively addresses the prediction model's deviation and variance through the use of a mutual information weighted ensemble learning model. The experimental results demonstrate that the proposed method is competitive compared to other comparison methods.

2. Theory and method

2.1. Functional mean variance estimation

$(x_1, y_1), \cdots, (x_n, y_n)$ is n independent random samples of (x, y), where $x \epsilon L_2(T)$ is a random function predictor and y is a response variable. In order to estimate the Effective Dimension Reduction (EDR) space, this paper uses the basis function expansion to project the function predictor $x(t)$ and the EDR direction β_j in the space spanned by the basis function[13].

$$x(t) \approx \sum_{i=1}^{k} x_i B_i(t), \qquad \beta_j(t) \approx \sum_{i=1}^{k} b_{ji} B_i(t) \tag{1}$$

Where k is the number of basis functions and $\beta_j(t)$ is a fixed basis function. In this algorithm, we first estimate the eigenvectors corresponding to the k largest eigenvalues of $\Gamma^{-1}E[\Gamma - 2var(x|y) + var(x|y)\Gamma^{-1}var(x|y)]$,where k is the dimension of EDR space E_k.The covariance operator of x ,$E(x|y)$ and $x|y$ can be estimated by the following three equations:

$$\hat{\Gamma} = \sum_{j=1}^{n} (x_i - \bar{x}) \otimes (x_i - \bar{x})/n \tag{2}$$

$$v\hat{a}r(E(x|y)) = \sum_{h=1}^{H} \bar{x}_h \otimes \bar{x}_h / H \tag{3}$$

$$v\hat{a}r(x|y) = \sum_{h=1}^{H} v\hat{a}r_h / H \tag{4}$$

where $\bar{x}$ is the average value of $x_1, \cdots, x_n$, and $\bar{x}_h$ is the sample average of the predictors in the hth slice. $v\hat{a}r_h = \sum_{j=1}^{c_h} (x_{(h,j)} - \bar{x}_h) \otimes (x_{(h,j)} - \bar{x}_h)$,$x_{(h,j)}$ is the jth observation value in the hth slice, and c_h is the number of observations in the hth slice. For functional data, even if the inverse matrix Γ^{-1} of the covariance operator Γ of x exists, it is usually an unbounded infinite dimensional operator. In the functional sufficient dimension reduction literature, people generally use the finite rank operator $\Gamma_k = \Pi_k \Gamma \Pi_k$ to replace Γ, where Π_k is an operator projected into the subspace S_k

covered by the first k basis functions. So even if Γ^{-1} is unbounded, the scope of $\Gamma_k{}^{-1}$ can be determined. The specific steps of functional mean variance estimation are as follows:

Step1: The basis functions $B_1(t),\cdots,B_k(t)$ are used to expand x_i and β_j. We can get the approximate value $x_i(t)=x_i^T B(t)$,where $x_i=(x_{i1},\cdots,x_{ik})^T$, $B(t)=(B_1(t),\cdots,B_k(t))^T$ .The coefficient matrix is $x_{n\times k}=(x_1^T,\cdots,x_n^T)^T$, $B_{k\times k}=(\langle B_i,B_j\rangle)_{i,j=1,\cdots,k}$

Step2: Cut y into H intervals $I_1,\cdots,I_H$. In each slice, the sample mean $\bar{x}_h$ and sample covariance $\hat{cov}_h$ of x_i are calculated.

Step3: Calculation

$$\hat{V}=\mathbf{B}^{-1}\left(v\hat{a}r(\mathbf{x})\right)^{-1}\mathbf{B}^{-1}[\mathbf{B}(v\hat{a}r(\mathbf{x}))\mathbf{B}-2\mathbf{B}\sum_{h=1}^{H}\frac{c_h}{n}c\hat{o}v_h\mathbf{B}$$
$$+\mathbf{B}\sum_{h=1}^{H}\frac{c_h}{n}c\hat{o}v_h v\hat{a}r(\mathbf{x})^{-1}c\hat{o}v_h\mathbf{B} \tag{5}$$

where c_h is the number of observations in the hth slice. Let the k largesteigenvalues of $\hat{V}$ be$\hat{\lambda}_1\geq\cdots\geq\hat{\lambda}_k$, and the corresponding eigenvectors are$\hat{\eta}_1,\cdots,\hat{\eta}_k$. Therefore, the estimated EDR direction is

$$\hat{\beta}_1(t)=\sum_{b=1}^{k}\hat{\eta}_{b1}B_b(t),\cdots,\hat{\beta}_k(t)=\sum_{b=1}^{k}\hat{\eta}_{bk}B_b(t) \tag{6}$$

Where $\hat{\eta}_1=(\hat{\eta}_{11},\cdots,\hat{\eta}_{k1})^T,\cdots,\hat{\eta}_k=(\hat{\eta}_{1k},\cdots,\hat{\eta}_{kk})^T$.

2.2. Ensemble learning based on mutual information weighting

Bagging proposed by Leo Breiman[14] in 1994, is the most famous representative of parallel ensemble learning. The idea of this algorithm is to let the learning algorithm train multiple rounds. The training set of each round is composed of M training samples randomly taken from the initial training set. An initial training sample can appear multiple times or no similar sampling in a round of training set. After training, a prediction function sequence can be obtained. When combining the predicted output, Bagging usually uses simple voting for classification tasks and simple averaging for regression tasks. The specific steps of the algorithm are as follows :

Step1： Randomly select m training samples from the original training samples.

Step2： The model is obtained by training the extracted m samples according to the given learner algorithm $f_k(x_i)$.

Step3： Put the training samples back. Repeat the above three steps for k times to obtain k model sets$\{f_1,f_2,\cdots,f_k\}$,The final prediction model is As shown in Fig.1

$$F(x_i)=\frac{1}{K}\sum_{k=1}^{K}f_k(x_i) \tag{7}$$

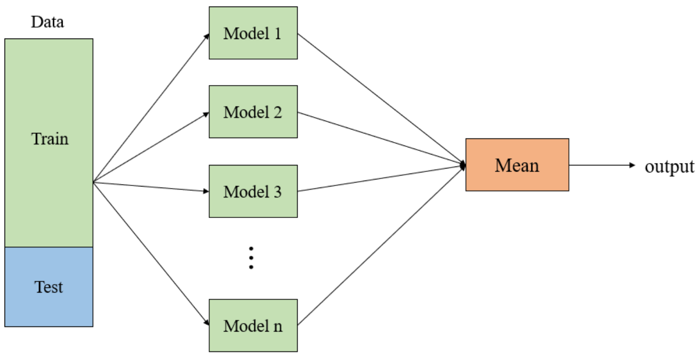

Figure 1. Bagging Algorithm

In order to further improve the prediction accuracy of Bagging algorithm. The performance of the model is improved by weighted combination. Mutual information is a measurement method used in signal science to measure the loss or difference before and after signal transmission. Considering that mutual information can measure the common information between two random variables, the stronger the correlation between the two variables, the greater the mutual information value ; if the two variables are independent of each other, the mutual information value is 0. This paper uses mutual information to determine the weight of a single learner. For the two continuous random variables $x=\{x_1, x_2, \cdots, x_d\}$ and $Y=\{y_1, y_2, \cdots, y_d\}$ in this paper, the calculation between mutual information is as follows [15]:

$$I(X,Y) = \int_x^X \int_y^X p(x,y) \lg \frac{P(u,v)}{p(u)p(v)} dudv \tag{8}$$

The resulting mutual information weighted combination model is:

$$F = \frac{\sum_{i=1}^{d} I(x_i,y_1) \times f_i}{\sum_{i=1}^{d} I(x_i,y_1)} \tag{9}$$

The projection data after functional dimension reduction is brought into the ensemble learning based on mutual information weighting to complete the modeling step. Specifically, first, we perform functional sufficient dimension reduction on functional data (here we use the grid retrieval and basis BIC criterion to determine the number of basis functions). Then, we use the bagging algorithm based on mutual information weighting to adaptively determine the best prediction model.

3. Data analysis

3.1. Data source and experimental description

The experimental data comes from http://lib.stat.cmu.edu/datasets/tecator/,The second experimental data comes from https://eigenvector.com/data/Corn/.As shown in Fig.2, the first data is the correlation between Waveleng and Absorbances of corn. The data set has 100 dimensions, namely X1-X100.The content of protein, fat and moisture was calculated by the correlation between 100 dimensions, and the data set contained a total of 216 samples. The second data is the correlation between Waveleng and Absorbances of meat. The data set has 700 dimensions, namely X1-X700. Oil, starch, protein and moisturede are calculated by the correlation between 700 dimensions. The data set contains a total of 81 samples.

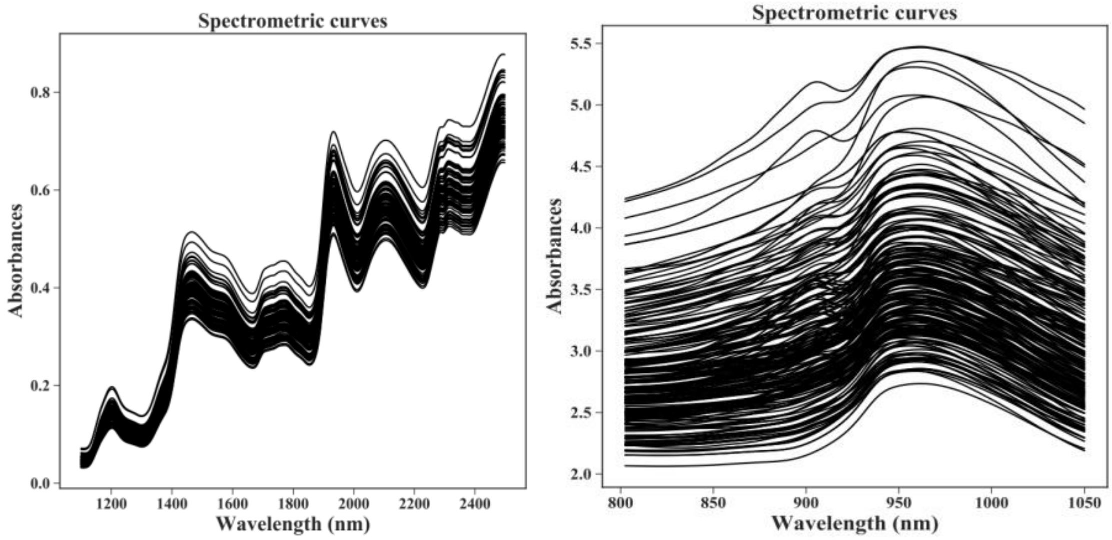

Figure 2. Spectrometric curve of corn and meat

Due to the small number of samples of experimental data, the experimental results obtained by using one data are accidental. Therefore, we use the numpy expansion library in python to expand the sample data, and randomly divide the training set and the test set 100 times. Each csv file contains 216 samples, that is, 100 * 216 experimental data are obtained. A large number of sample data ensure the universality of the experimental results and are persuasive. The experimental environment is specifically computer programming, the operating system is Windows 11, the experimental process uses the programming language R language and python, and the software is Rstudio and Pycharm.

We use the basis function expansion method to reduce the dimension of high-dimensional data, avoiding the information loss caused by the dimension reduction of high-dimensional data. Then, the weight of the bagging algorithm is determined based on the functional average variance estimation and mutual information weighting. The bagging algorithm is used to regress the y value of the data, and compared with the common traditional regression methods, such as Linear regression model (linear), decision tree regression model (regression tree), random forest model (random forest), LASSO model ridge regression model (LASSO). The MSE and MRE of the prediction results of KNN model (KNN) and SVM regression model (SVM) were compared to evaluate the accuracy of the regression model.

3.2. comparing the results

By comparing the prediction results of y with the traditional regression model, we find that the prediction results of the model have higher accuracy, that is, there are smaller MSE and MRE.We draw the box plot of the mean square error and absolute error of different regression methods. The comparison results show that the model does have smaller mean square error and absolute error. The specific results are shown in Table 1-2 and Figure 3-4: MSE_m and MRE_m respectively represent the mean square error and the mean absolute error of 100 experiments. MSE_s and MRE_s represent the standard deviation of mean square error and absolute error of 100 experiments respectively.

$$\mathrm{SSE} = \sum_{i=1}^{m} w_i (y_i - \hat{y}_i)^2 \tag{10}$$

$$\mathrm{MSE} = \frac{\mathrm{MSE}}{n} = \frac{1}{n}\sum_{i=1}^{m} w_i (y_i - \hat{y}_i)^2 \tag{11}$$

$$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{m} |y_i - \hat{y}_i| / y_i \tag{12}$$

Table 1. Mean, std of SME and MRE of different regression methods in corn dataset

	SME_m	SRE_m	SME_s	SRE_s
		Fat		
TREE	9.382276923	13.75468136	1.122030524	1.408420527
KNN	9.062664615	11.60885174	0.811398665	0.921658645
SVM	9.193574820	11.86474408	0.884062258	1.131693055
RF	8.218859692	10.85299716	0.796356821	1.036929840
RIDGE	9.080360232	11.33607269	0.734919577	0.967182414
LASSO	9.082262391	11.33560922	0.734688974	0.966232269
LINEAR	9.080334152	11.33611856	0.734929131	0.967215276
FSDR_MLB	3.093173382	3.863281665	0.762453877	0.895426077
		Protein		
TREE	2.362707692	3.323414857	0.282015663	0.336465562
KNN	2.380058462	2.924126057	0.188079710	0.213228878
SVM	2.391398056	3.012934474	0.176699976	0.212965911
RF	2.094560462	2.645770404	0.181194695	0.226583833
RIDGE	2.234488987	2.735218851	0.171091027	0.186804086
LASSO	2.236858747	2.735377143	0.170703626	0.186218602
LINEAR	2.234460707	2.735221485	0.171095593	0.186813746
FSDR_MLB	1.045695369	1.297564172	0.122056348	0.153956584
		Water		
TREE	7.287176923	10.60838618	0.88964929	1.117573577
KNN	7.083903077	8.903436605	0.555016408	0.629520738
SVM	6.996452627	9.221478066	0.545662238	0.697709704
RF	6.383429846	8.322794335	0.556343140	0.735848485
RIDGE	7.156623109	8.751980361	0.414913850	0.560053532
LASSO	7.159447854	8.751863794	0.414758387	0.559036483
LINEAR	7.156587844	8.752009347	0.414913768	0.560075050
FSDR_MLB	2.490226477	3.042310724	0.410944214	0.502051824

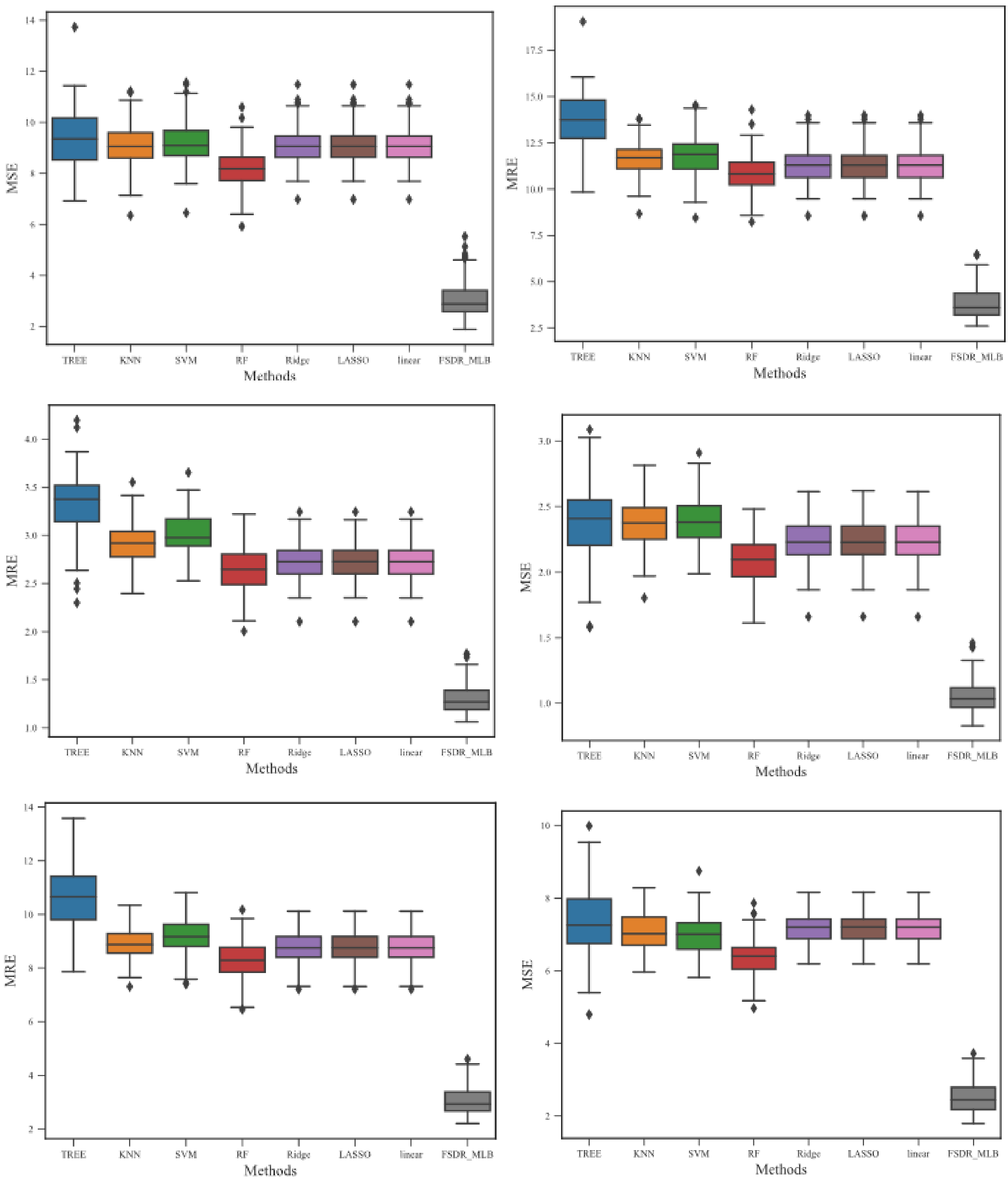

Figure 3. Different regression methods were used to regress MSE and MRE box plots of corn data set fat, protein, water.

Firstly, according to the regression of Fat, protein and water by different methods in the above table, we can see that the mean of MSE and MRE is higher, indicating that the regression results of these methods have a larger error, and MRE _ s, MSE _ s also have higher values, indicating that the MRE and MSE of the results obtained by several traditional regression methods are relatively discrete.At the same time, according to the box plot results, we can intuitively see that the functional data regression model based on functional average variance estimation and mutual information weighted ensemble learning has smaller mean square error and absolute error than other models, and the regression results are better.

On the one hand, the proposed method uses the functional mean variance estimation to reduce the data dimension, and extracts the potential function information of the original data by projecting the random function into the function space spanned by the finite dimensional basis function, and this extraction method does not lose the effective information for the response variable. On the other hand, considering that the number of sufficient dimension reduction sub-directions and the degree of interpretation of response variables in each direction are unknown, we propose to use the model averaging method to adaptively solve the over-fitting and under-fitting problems of the prediction model. In addition, due to the different importance of each base model, this paper proposes to use mutual information weighting to measure the importance of each sub-model. Therefore, by weighting the weight of the bagging algorithm with a weighted model, a prediction result with smaller error can be obtained.

Table 2. Mean, std of SME and MRE of different regression methods in corn dataset

		Moisture		
	SME_m	**SRE_m**	**SME_s**	**SRE_s**
TREE	0.279533500	0.364576269	0.052114700	0.068127391
KNN	0.256475500	0.317332308	0.041485901	0.049219941
SVM	0.249568548	0.300595382	0.035636451	0.042955780
RF	0.256806680	0.324889181	0.040977491	0.054037678
RIDGE	0.242396896	0.299203242	0.037394723	0.043901222
LASSO	0.241428001	0.299171184	0.038376721	0.044526879
LINEAR	0.242447971	0.299252937	0.037351124	0.043861074
FSDR_MLB	0.179586544	0.223765844	0.036498637	0.041515141
		Oil		
TREE	0.165789000	0.207943258	0.023498569	0.027656220
KNN	0.138820600	0.172057551	0.021264556	0.028009265
SVM	0.138432003	0.169996876	0.018464892	0.022813216
RF	0.144527890	0.179707948	0.022350777	0.026107466
RIDGE	0.139432744	0.169339110	0.018767932	0.022882909
LASSO	0.139722304	0.170385225	0.018568612	0.021190459
LINEAR	0.139426658	0.169331789	0.018769471	0.022890317
FSDR_MLB	0.102007210	0.121901003	0.014919951	0.015614149
		Protein		
TREE	0.551098000	0.686693750	0.079190526	0.076351201
KNN	0.431970400	0.518026613	0.061868306	0.067151186
SVM	0.382178307	0.483401934	0.059752536	0.064509722
RF	0.480607220	0.591103901	0.070724426	0.070608088
RIDGE	0.388895350	0.476697627	0.052505571	0.053256172
LASSO	0.388576333	0.476443965	0.052482113	0.053057348
LINEAR	0.388898646	0.476700814	0.052505773	0.053257262
FSDR_MLB	0.227181450	0.275672033	0.039096828	0.045969025
		Starch		
TREE	0.942913500	1.145373297	0.130303341	0.138290460
KNN	0.717663700	0.894790821	0.104857163	0.098935075
SVM	0.660098247	0.828313115	0.103278033	0.107380886
RF	0.812548055	0.993114841	0.113128494	0.113102997
RIDGE	0.688735405	0.842295879	0.102311786	0.108748815
LASSO	0.687802037	0.841728539	0.102031561	0.108347475
LINEAR	0.688744641	0.842306300	0.102312374	0.108750804
FSDR_MLB	0.425080797	0.538564307	0.084169991	0.093346450

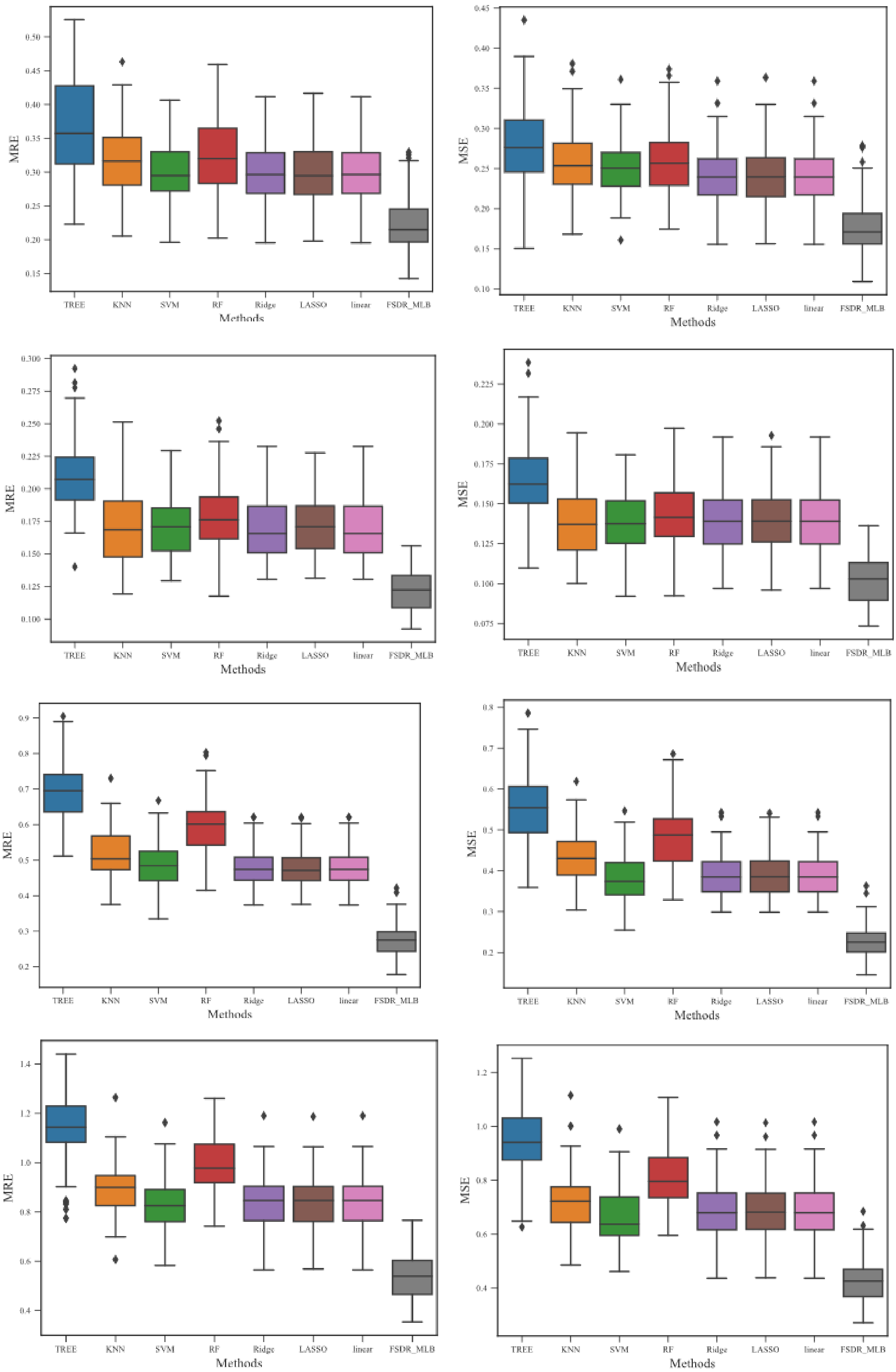

Figure 4. Different regression methods were used to regress MSE and MRE box plots of meat data set fat, protein, water.

4. Conclusions

For functional data, this paper proposes a new regression method. This method includes the principal component basis expansion of functional data, the sufficient dimension reduction method of slice average variance estimation and the machine learning model weighted by mutual information. On the premise that the intrinsic function information of the original data is fully retained, the problem of data dimension disaster is solved, and then the mutual information weighting is applied to the Bagging algorithm, so that the importance of the response variables in each direction of the full dimension reduction is reflected, and the over-fitting and under-fitting problems of the model are effectively processed.

For future research directions, first of all, about the weighting criteria of mutual information, some literature proposes that the mutual information obtained for the same and different is generally different. It is not comprehensive enough to determine the amount of information containing the overall quality Y information only according to the size of the mutual information value with a certain one. Each mutual information value can be integrated into a weighted fusion value to make the model cover more information. Secondly, there are many ways to average the final model. Heuristic learning uses the least relevant information, past results and experience to generate feasible and practical solutions for the problem in a reasonable time. This strategy focuses on providing fast results with acceptable accuracy range. If heuristic learning is used with optimization algorithms, the average model with higher accuracy can be further determined.

References

[1] Liu Zhuang. Functional data classification method and its application [D]. Hefei University of Technology, 2020.

[2] DING Hui, XU Wenchao, et al. Review of Regression Analysis for Function Data [J]. School of Statistics, East China Normal University, College of Economics, et al., 2018,34(06):630-654.

[3] Li Xingxing. Comparisons of generalized linear models [D]. Yangzhou University, 2017.

[4] CHEN X, CUI Y. Quasi-likelihood adaptive Lasso estimators for high-dimensional generalized linear models [J]. Journal of Shaanxi Normal University (Natural Science Edition),2019,47(2):1-9.

[5] Chanjuan Li. A Research of Key Technology of dimension Reduction of High Dimensional Data [D]. School of Information and Software Engineering, 2017.

[6] Marukatat, S. Tutorial on PCA and approximate PCA and approximate kernel PCA [J]. Artificial Intelligence Review,2022.

[7] Shigemoto, H; Morimoto, T. Forecasting High-Dimensional Covariance Matrices Using High-Dimensional Principal Component Analysis [J]. Axioms, 2022,11(12):692.

[8] Min Cai. Statistical computation and theory research of sufficient dimension [D]. East China Normal University, 2022.

[9] YUAN Xiao-hui, CAO Ru-ya, JIN Wan-lin. Parameter estimation and variable selection for function partially linear models [J]. Journal of Jilin University (Natural Science Edition), 2022,43(02):48-57.

[10] Tianli Xiao. The Significance Test and Effect Analysis of Functional Parameters [D]. Zhengzhou University, 2018.

[11] Guochang Wang, Nan Lin, Baoxue Zhang. Dimension reduction in functional regression using mixed data canonical correlation analysis [J]. Statistics And Its Interface, 2013, 6(2): 187-196.

[12] Banner KM, Higgs MD. Considerations for assessing model averaging of regression coefficients [J]. Ecological Applications, 2017,27(1):79-93.

[13] Guochang Wang, Yan Zhou, Xiang-Nan Feng, Baoxue Zhang. The hybrid method of FSIR and FSAVE for functional effective dimension reduction [J]. Computational Statistics and Data Analysis, 2015,91(1): 64-77.

[14] Li Yaqin, Yang Huizhong. Ensemble modeling method based on Bagging algorithm and Gaussian process [J]. Journal of Southeast University (Natural Science Edition), 2011,41(05):93-96.

[15] HU Jian, SU Yongdong, et.al. Intrusion detection method based on ensemble transfer learning via weighted mutual information [J]. Journal of Computer Application, 2019, 39(11): 3310-3315.