

High-Frequency Time Series Data Regression Based on the Functional Bayesian Model Averaging

Jiarui Cui^{#,1,*}, Shuo Yang^{#,2}, Dinghui Chen^{#,3}

¹ School of Mathematics and Physics, University of Science and Technology Beijing, Beijing 100083

² School of Mathematics and Information Science, Yantai University, Yantai, Shandong 264005

³ Digital Trade Academy, Zhejiang Yuexiu Foreign Chinese College, Shaoxing, Zhejiang 312030

* Corresponding Author Email: cjr905931051@163.com

[#]These authors contributed equally.

Abstract. Regression analysis of high-frequency time series data has been an important problem in the field of statistical learning. In this paper, we propose a novel approach based on multiple basis function expansions and Bayesian model averaging when the predictor variable is a high-frequency time series and the response variable is continuous scalar data. On the one hand, the proposed method avoids the curse of dimensionality and extracts the functional information of the original sequence by transforming the discrete data into continuous functions in a function space consisting of some fixed basis functions. On the other hand, for the choice of basis functions, this paper proposes to adaptively determine the optimal solution using Bayesian model averaging, effectively balancing the variance and bias of the predictive model. The real data analysis shows that the proposed method has smaller mean squared error and absolute error, and is robust compared to other methods. Finally, the proposed method can be further extended to time series data applications such as weather forecasting and stock price prediction.

Keywords: Time series data, Regression, Model averaging, Functional data analysis.

1. Introduction

High-dimensional time series data reflect the state or degree of a phenomenon or object changing over time, such as financial stock prices, real-time temperature, ECG curves, etc. In this paper, we consider the case where the predictor variable is a time series and the response variable is a continuous variable. Specifically, the infrared spectrometer extracts the wavelength data of the absorption spectrum of corn or meat over time, which are high-frequency time series data. The task of this paper is to extract and utilize the latent information of such high-frequency time series data to predict the nutritional value of corn and meat, such as protein, moisture, starch and oil content.

Given this data background, a multivariate linear model can be used for analysis. However, this inevitably suffers from the curse of dimensionality. What's more, there are dependencies between different time points in time series data, which makes it impossible for classical linear regression models to identify. At present, there are roughly two mainstream processing methods. The first is regularization techniques such as ridge regression and LASSO regression. Specifically, there is a prediction model built through LASSO regression to analyze financial stock price trends. Guo (2021) [1]. There is also a LASSO-CDRD covariance matrix prediction model based on high-frequency financial data. Liu (2022) [2]. And a penalty-likelihood-ratio-based EWMA quality monitoring model for high-dimensional spatial correlation processes. Zhang (2023) [3]. The second method is dimensionality reduction techniques, such as PCA (Principal Component Analysis), PLS (Partial Least Squares) and other methods. Specifically, for example, in the prediction of inactivated *Bacillus subtilis* content based on hyperspectral imaging, PLS is used to reduce the dimensionality of pork characteristic wavelengths. Zhang (2023) [4] Research on Hyperspectral Prediction of Bloodstain Aging Based on Partial Least Squares. Sun (2022) [5] and the study of multispectral dimensionality reduction algorithms based on principal component analysis (PCA) and second order with polynomial

regression. Cao (2022) [6] These methods allow regression prediction in a suitable space, thus produce more efficient predictive values. For nonlinear relationships, decision trees, random forests or kernel PCA can also be used for processing, which can effectively improve the prediction accuracy. For example, optimization of random forest hyperspectral inversion of soil selenium content based on Seagull algorithm. Xie (2022) [7] or reconstruct asphalt pavement surface texture and predict construction depth through decision tree. Yang (2023) [8]. However, these methods cannot identify functional information from high frequency time series data.

In this paper, we propose a regression model based on smooth basis functions and Bayesian model averaging. Specifically, the innovative of the proposed method is that, on the one hand, high-frequency time-series data is naturally treated as functional data, and the method of basis function expansion is used to extract potential function information, and the basis function expansion coefficients are used to replace the original sequence for prediction to achieve dimensionality reduction. On the other hand, considering the unknown real function space, we propose to mix common basis functions such as B-spline basis functions, functional principal component basis functions, and Fourier basis functions and combine them with Bayesian model averaging. By using model averaging instead of model selection, the problem of overfitting and underfitting of predictive models is effectively balanced. Finally, the results of the actual data analysis show that the proposed method is competitive with the comparison methods such as ridge regression and LASSO.

2. Theory and Methods

2.1. Expansion with Basis Functions

High-dimensional time series data have weak dependencies between them. In the case of discrete data, it is not suitable to directly process the data as multivariate data. Therefore, this article will transform the discrete data into a functional form by using a linear combination of basis functions to convert the discrete values into a continuous and smooth function. Assuming $x^i(t)(i=1,2,...,n)$ This is the function object to be studied, and we can obtain $x^i(t)$ and T_i observations $y^i = (y_1^i, y_2^i, ..., y_{T_i}^i)'$ [9]. Specifically:

$$y_j^i = x^i(t_j^i) + \varepsilon^i(t_j^i), (i=1,2,...,T_i) \quad (1)$$

To estimate $x^i(t)$, we assume that a set of basis functions is $\phi(t) = \{\phi_1(t), \phi_2(t), ..., \phi_K(t)\}'$, and we expressed $x^i(t)$ as a linear combination of these basis functions:

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

Its matrix form is:

$$x^i(t) = c^i \phi(t), c^i = (c_1^i, c_2^i, ..., c_K^i)' \quad (3)$$

Finally, the penalty least squares method is used to obtain the vector of basis expansion $x^i(t)$ coefficients of the function c^i , which can be expressed by the following formula:

$$c^i = \arg \min_{c^i} \sum_{j=1}^{T_i} [(y_j^i - c^i \phi(t_j^i))^2 + \lambda \|c^i\|_1] \quad (4)$$

Both general polynomial functions and step functions can be used for basis function expansion, or Fourier series and wavelet transform can be used for basis function expansion. Regression splines, which are an extension of polynomial regression and step functions, can also be used for basis function expansion.

2.2. Bayesian Model Averaging

Bayesian Model Averaging (BMA) is a statistical method that uses a combination of multiple models for probabilistic prediction. BMA uses Bayesian formula to combine the prior distribution of location parameters with prior information and the likelihood function that describes the new information of the object to obtain the posterior distribution of the object. Based on the posterior distribution, BMA performs statistical inference and prediction of the object, thereby improving the understanding of the object. BMA estimates all possible combinations of the latent variable X and establishes the weight of each model through weighted averaging after selecting the necessary models for the combination. Taking into account the uncertainty of the models themselves, BMA can generate a probability density function with predictive effect to solve the problem. The predictive probability density function of BMA is:

$$p(y | f_1, \dots, f_k) = \sum_{k=1}^K w_k p_k(y | f_k) \quad (5)$$

Where y is one of the predicted variables, f_k representing the original predicted value of the k -th set member $p_k(y | f_k)$ based on f_k . The predicted probability density function, and K represents the number of set members. It is the posterior probability when the k -th member is w_k the best prediction, and the sum of the posterior probabilities is 1, which can also be called the weight[10]. As indicated (5) by the equation, y the posterior distribution of $P(f_k | Z)$ is actually a weighted average of the posterior distributions of all models, with the posterior probabilities serving as weights. In particular, assuming f_k is the optimal model, its probability distribution is:

$$P(y | f_k, Z) = \int P(f_k | \beta, f_k, Z) P(y | f_k, Z) \quad (6)$$

where $P(y | f_k, Z)$ is the posterior distribution given the survey data Z and model conditions f_k ; $P(f_k, Z)$ is the probability of the f_k optimal model under the Z data conditions. For there $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_n)$ gression coefficient vector of the model f_k .

2.3. Bayesian averaging model based on basis function expansion

The methods based on ridge regression or Lasso regression have the drawback of ignoring the existence of weak dependence between time-series type data, which reduces the prediction accuracy to some extent. We can regard the discrete function-type data as the realization of discrete points of a function, re-functionalize it and reconstruct the function; we can project the data into the function space to get the coordinates of the discrete data, and use the projected coordinates in the function space obtained from the time-series data instead of the original time-series data, which can both consider the function information and avoid the dimensional disaster. Let the number of dimensions of the time-series data be N . Since there are many kinds of basis functions, using a single model selection method is easy to make the prediction model fall into the over-fitting or under-fitting situation, You can specify specific classes of better fitting basis functions constructed as a set $\Omega = \{\alpha_1(x), \dots, \alpha_M(x)\}$ Construct function spaces separately using each type of basis function in Ω . Project the time-series data into different function spaces separately. Generate the projection coordinates of group M : $A = \{\varphi_1, \varphi_2, \dots, \varphi_M\}$. Combining the M sets of projection coordinates into an $N \times M$ -dimensional coordinate vector $X_{N \times M}$, which is then averaged over a Bayesian model.

$$p(y | \varphi_1, \varphi_2, \dots, \varphi_M) = \sum_{i=1}^M w_i p_i(y | \varphi_i) \quad (7)$$

where y is the predictor variable, w_i is the weight of each model, and $\sum_{i=1}^M w_i = 1$. Then, the weights of each model are obtained and multiple models are linearly combined to obtain the model with the better fit, resulting in the following predictions.

$$y = \sum_{i=1}^M w_i \varphi_i \quad (8)$$

The specific implementation steps of the appealed method are shown in Figure 1 below:

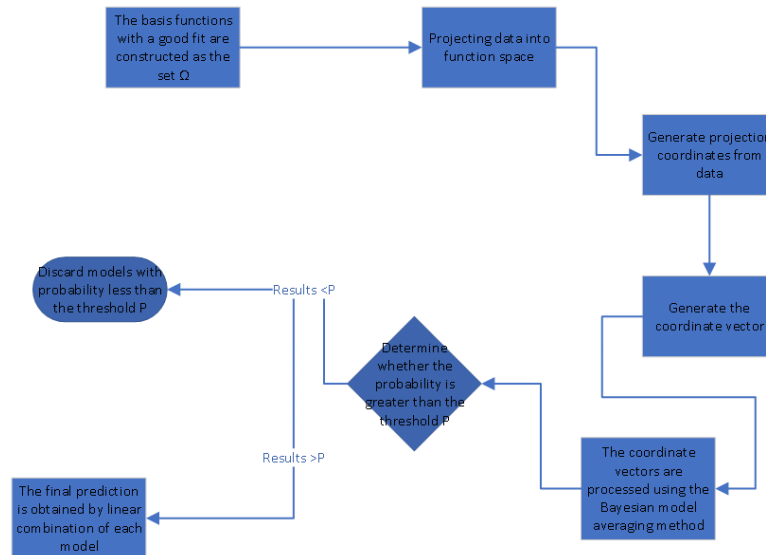


Figure 1. Algorithm flow

3. Data analysis

3.1. Data Source and Experimental Description

The dataset used in this article comes from the websites <http://lib.stat.cmu.edu/datasets/tecator/> and <https://eigenvector.com/data/Corn/>. The first dataset consists of 215 meat samples, each with a dimension of 100, measuring their fat content, moisture, and protein values. The second dataset consists of 80 corn samples, each with a data dimension of 700, measured using three different near-infrared spectrometers to obtain their moisture, oil, protein, and starch values. Five-fold cross-validation was used to split the training and testing sets for both datasets, and then regression models were trained on them. as shown in Figure 2.

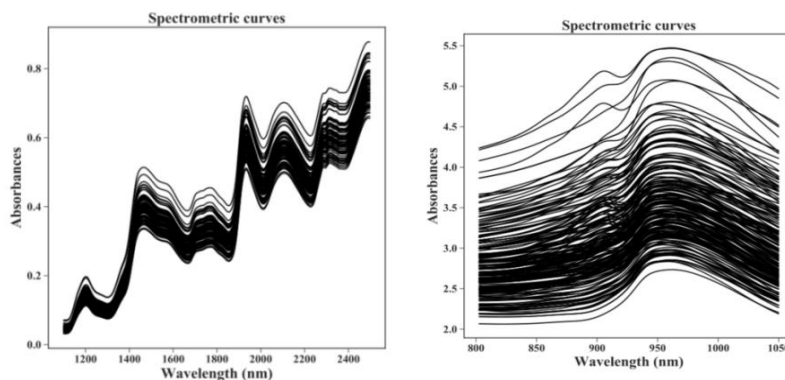


Figure 2. Infrared spectrum wavelengths of meat (right) and corn (left)

The experimental environment was R language and Python. In order to compare the test results of different models, mean squared error (MSE) and absolute error were selected as evaluation indicators.

MSE is the expected value of the square of the difference between the parameter estimate and the true value, which can evaluate the degree of data variation. The smaller the MSE value, the better the precision of the predictive model to describe experimental data. The mean absolute error better reflects the actual situation of the prediction value error, with the specific formula as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (9)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (10)$$

where $\hat{y}_i$ is the predicted value, and y_i is the true value. In terms of selecting the basis functions, functional principal component bases and Fourier bases were chosen, Linear regression, Ridge regression, and LASSO Regression were used to horizontally compare the mean and standard deviation of MSE and MAE of seven models. Our method is the fusion of functional principal component bases and Fourier bases using Bayesian averaging (FBMA). The other six comparative experiments are Fourier-based LASSO regression (FLASSO), Fourier-based ridge regression (FRidge), Fourier-based linear regression (Flinear), functional principal component bases-based LASSO regression (FPCA_LASSO), functional principal component bases-based ridge regression (FPCA_Ridge), and functional principal component bases-based linear regression (FPCA_linear). The experimental results are shown in Tables 1-7:

Where “aE+b” is for scientific counting.

Table 1. tecator_Fat errors under diffierent models

Methods	MRE		MSE	
	mean	std	mean	std
FBMA	2.34E+00	1.88E-01	8.97E+00	1.93E+00
FLASSO	2.65E+00	2.17E-01	1.12E+01	2.04E+00
FRidge	2.75E+00	2.24E-01	1.15E+01	1.91E+00
FPCA_LASSO	3.09E+00	2.43E-01	1.51E+01	2.58E+00
FPCA_Ridge	3.19E+00	2.49E-01	1.53E+01	2.36E+00
Flinear	2.75E+00	2.24E-01	1.15E+01	1.91E+00
FPCA_linear	3.19E+00	2.49E-01	1.53E+01	2.36E+00

Table 2. tecator_Water errors under diffierent models

Methods	MRE		MSE	
	mean	std	mean	std
FBMA	2.34E+00	1.88E-01	8.97E+00	1.93E+00
FLASSO	2.65E+00	2.17E-01	1.12E+01	2.04E+00
FRidge	2.75E+00	2.24E-01	1.15E+01	1.91E+00
FPCA_LASSO	3.09E+00	2.43E-01	1.51E+01	2.58E+00
FPCA_Ridge	3.19E+00	2.49E-01	1.53E+01	2.36E+00
Flinear	2.75E+00	2.24E-01	1.15E+01	1.91E+00
FPCA_linear	3.19E+00	2.49E-01	1.53E+01	2.36E+00

In the predictive experiments on the fat, moisture, and protein content of meat, the proposed method had a mean MSE and MAE that were both lower than the other six comparison methods. In the predictive experiments on the nutritional elements of corn, such as oil content, the proposed method had a better predictive performance in terms of mean MSE compared to methods such as FLASSO. For the problem of predicting high-dimensional function data, the Bayesian model averaging method was adopted, which trained the model by inputting high-dimensional function

information and weighted the final result based on the maximum posterior probability criterion using Bayesian theory, resulting in a better prediction performance.

Table 3. tecator_Protein errors under different models

Methods	MRE		MSE	
	mean	std	mean	std
FBMA	8.27E-01	1.30E-01	7.27E-01	5.51E-02
FLASSO	1.32E+00	1.99E-01	9.21E-01	7.31E-02
FRidge	1.43E+00	2.63E-01	9.52E-01	8.63E-02
FPCA_LASSO	2.43E+00	4.97E-01	1.20E+00	1.09E-01
FPCA_Ridge	2.44E+00	4.92E-01	1.20E+00	1.11E-01
Flinear	1.43E+00	2.63E-01	9.52E-01	8.63E-02
FPCA_linear	2.44E+00	4.92E-01	1.20E+00	1.11E-01

Table 4. m5spec_Protein errors under different models

Methods	MRE		MSE	
	mean	std	mean	std
FBMA	2.21E-07	6.05E-08	3.67E-04	5.38E-05
FLASSO	1.21E-04	5.28E-05	8.51E-03	1.68E-03
FRidge	8.33E-06	3.58E-06	2.27E-03	4.39E-04
FPCA_LASSO	1.23E-04	5.32E-05	8.58E-03	1.69E-03
FPCA_Ridge	3.65E-05	1.51E-05	4.64E-03	8.66E-04
Flinear	5.04E-07	1.96E-07	5.69E-04	9.54E-05
FPCA_linear	1.66E-06	6.11E-07	1.00E-03	1.69E-04

Table 5. m5spec_Oil errors under different models

Methods	MRE		MSE	
	mean	std	mean	std
FBMA	2.33E-07	6.14E-08	3.81E-04	5.90E-05
FLASSO	1.26E-04	5.96E-05	8.71E-03	1.79E-03
FRidge	7.78E-06	3.51E-06	2.20E-03	4.47E-04
FPCA_LASSO	1.28E-04	6.01E-05	8.79E-03	1.80E-03
FPCA_Ridge	3.83E-05	1.83E-05	4.77E-03	9.45E-04
Flinear	4.75E-07	1.78E-07	5.45E-04	9.86E-05
FPCA_linear	1.75E-06	7.23E-07	1.05E-03	1.74E-04

Table 6. m5spec_Moisture errors under different models

Methods	MRE		MSE	
	mean	std	mean	std
FBMA	2.69E-07	6.68E-08	4.28E-04	5.91E-05
FLASSO	1.32E-04	5.75E-05	8.75E-03	1.71E-03
FRidge	8.58E-06	4.22E-06	2.17E-03	5.05E-04
FPCA_LASSO	1.34E-04	5.80E-05	8.82E-03	1.71E-03
FPCA_Ridge	4.08E-05	1.80E-05	4.81E-03	8.99E-04
Flinear	4.58E-07	1.85E-07	5.31E-04	1.06E-04
FPCA_linear	1.88E-06	7.43E-07	1.06E-03	1.81E-04

The reasons for this were mainly due to the following aspects: firstly, by treating time series data as functional data, the proposed method effectively avoided information redundancy and extracted functional information. Secondly, because the prediction performance of the basis function expansion

method differs depending on the choice of basis functions, we proposed using multiple basis functions to functionize the original time series and used Bayesian model averaging to construct an ensemble learning model based on the submodel being a generalized linear model. This method can adaptively search for the optimal model, replacing the model selection method and effectively balancing the bias and variance of the predictive model. The proposed method comprehensively considered information feature extraction and model construction, thereby improving the overall accuracy of the predictive model.

Table 7. m5spec_Starch errors under different models

Methods	MRE		MSE	
	mean	std	mean	std
FBMA	2.41E-07	7.08E-08	3.81E-04	6.51E-05
FLASSO	1.23E-04	5.64E-05	8.58E-03	1.73E-03
FRidge	8.93E-06	5.11E-06	2.20E-03	5.68E-04
FPCA_LASSO	1.25E-04	5.68E-05	8.66E-03	1.74E-03
FPCA_Ridge	3.77E-05	1.73E-05	4.72E-03	9.24E-04
Flinear	4.76E-07	1.93E-07	5.34E-04	1.05E-04
FPCA_linear	1.73E-06	7.38E-07	1.02E-03	1.89E-04

4. Conclusion and Outlook

For the case where the predictive variable is a high-frequency time series and the corresponding variable is a continuous scalar data, this paper proposes a method based on multiple functional basis function expansions and Bayesian model averaging. Compared to LASSO regression and ridge regression, the method preserves the functional information of the original time series through basis function expansion. The Bayesian model averaging method not only solves the problem of selecting the type of basis functions but also effectively balances the overfitting and underfitting problems of the model. For future research directions, on the one hand, if the base model in Bayesian model averaging is a generalized linear model, using model-free weighted Bagging methods may be able to handle more complex data situations. On the other hand, the basis function expansion methods used in this paper are all semi-supervised learning techniques, and higher-level basis function expansion methods should be related to the response variable. Therefore, it is possible to try to further improve the proposed method by using supervised basis expansion methods such as functional dimensionality reduction techniques. Finally, the proposed method is not limited to predicting crop nutritional value and can be further applied to time-series data applications such as stock price prediction and weather forecasting.

References

- [1] Guo Z C. Stock Price Prediction and Analysis based on LASSO Method and Neural network [D]. Lanzhou university, 2021. DOI: 10.27204 /, dc nki. Glzhu. 2021.000313.
- [2] Liu Guangying, Bao Yueyan, Lin Jinguan. LASSO-CDRD Covariance Matrix Prediction Model Based on financial high frequency data [J]. Statistical Research, 2022(009):039.
- [3] Zhang Shuai, Yang Jianfeng, Liu Yumin, et al. EWMA Quality Monitoring Model of High Dimensional Spatial Correlation Process Based on Penalty Likelihood Ratio [J]. Operations Research and Management, 2022, 31(9):7.
- [4] Zhang Zehua, Liu Xiaohua, LAN Weijie, et al. Prediction of thermotrichothrix content in unpackaged and PE packaged chilled pork ribs based on hyperspectral imaging [J/OL]. Food and fermentation industry: 1-13 [2023-03-15].
- [5] Sun Wei, Liu Huize, Liu Jinkun, et al. Hyperspectral prediction of blood stain obsolescence based on Partial least square method [J]. Applied Chemical Industry, 2022(001):051.

- [6] Cao Qian, Xiao Ying. Research on Multi-spectral dimension Reduction Algorithm Based on Principal component Analysis and second-order polynomial regression [J]. Digital Printing, 2022(6):9.
- [7] Xie Peng, Wang Zhenghai, Xiao Bei, et al. Hyperspectral inversion of soil selenium content in random forest based on Gull algorithm [J]. Laser & Optoelectronics Progress, 2023, 60(17).
- [8] Yang Enhui, Chen Qiang, Li Jie, et al. Depth prediction model of surface texture reconstruction and construction of asphalt pavement [J/OL]. China journal of highway: 1-12 [2023-03-15].
- [9] Lu, Lele, Wang, Zhen, Zhang, Xiongqing, et al. Construction of a single wood diameter at breast height model for cedar based on Bayesian model averaging and stepwise regression[J]. Forestry Science, 2021, 57(9):11.
- [10] Chen Xiaofeng, Yin Ruifei. A functional data clustering method based on basis function expansion[J]. Statistics and Decision Making, 2009(19):3.