Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070110
ISSN 2371-8838 Vol. 7 Num. 1

Time series regression based on Bayesian model averaging and principal component analysis

Jiayi Lu

School of Urban, Xi'an Polytechnic University, Xi'an, Shaanxi, 710600, China

Keywords: Time series data, high-dimensional problem, PCA, model averaging

Abstract: This paper proposed an adaptive prediction model for high-dimensional time series data based on model averaging method and principal component analysis. Specifically, this paper considers the case where the response variable is a scalar and the predictor variable is a time series. Firstly, the high-dimensional time series data is extracted information by principal component analysis. Secondly, the Bayesian model averaging method is used to perform the forecast task based on the principal component projection matrix. The proposed method can effectively deal with the unsupervised nature of PCA and avoid the problem of selecting the number of PCA. It is demonstrated that the proposed method is competitive compared with the lasso regression and the ridge regression by real data analyses.

1. Introduction

In the era of rapid development of intelligence, people are confronted with complex types of data and huge information. The objects to be predicted are often affected by multiple related factors, and the dimensional space number of data indicator variables to be processed is usually one or more orders of magnitude of the sample number, which is called high-dimensional data. Therefore, research on feature extraction and prediction of high-dimensional data can make a great contribution to information processing and quantitative decision-making in various fields. Zhao Xun[1]and others used the principal component analysis and neural network algorithm to deal with high-dimensional data to predict the related parameters of residents' consumption and the relationship between the per capita consumption spending next year; Zhang He[2] et al. based on the Lasso regression model, predicted and analyzed the marine economic industry of Qingdao with 20 characteristic variables. Shi Yang [3] predicted soil composition based on neural network and the partial least squares method for high-dimensional data containing spectral absorbance of multiple wavelengths. Meng Qinglong et al. [4] performed principal component regression on spectral data of apples to predict the soluble solid content of apples. In this paper, high-dimensional time series data about meat content are used to carry out the experiment of the model. The predictive variables in the data record of each meat sample are the 100 wavelengths of screening, and these 100 variable indicators comprehensively reflect the organic matter content in meat. The time series data of meat are presented in the form of waves in Figure 1.

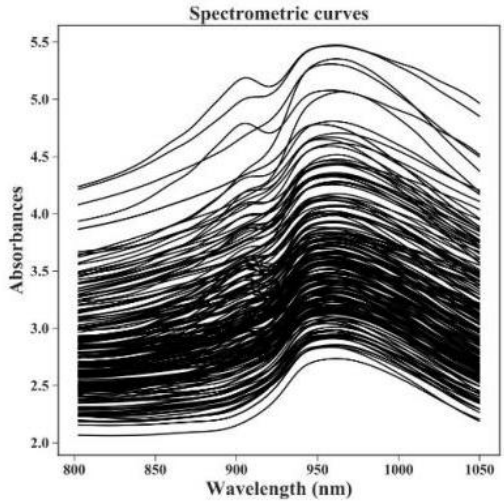

Figure 1: Meat spectrometric curves

Each line represents a meat sample and contains 100 absorbance degrees on each line, that is, the original dataset contains 100 dimensions. In the prediction of high-dimensional data, the principal component regression is often used in model building, and the contribution rate of variance and the magnitude of eigenvalues were used as the basis for selecting principal components, so it is easy to make the information with a large contribution rate in the front cover the information in the back. Moreover, the principal component regression algorithm carries out principal component analysis and multiple linear regression step by step. In the process of processing, the relationship between the selected principal components and the response variables is uncertain, so the established model may ignore the principal components strongly related to the response variables, thus making the prediction effect of the model poor. Some scholars have improved the principal component regression, which has significantly improved the selection of variables and the accuracy of prediction. Shi Yang [3] used partial least squares and stepwise regression models to predict the spectral data of soil composition and realized the screening of independent variables to obtain the optimal set of explanatory variables. Zhu Hailong [5] et al. proposed to analyze the influencing factors of Anhui provincial financial revenue through ridge regression and Lasso regression. Xu Yunjuan[6] et al. conducted Lasso dimensionality reduction of principal components based on variable clustering, and the accuracy of variable selection was improved. Li Yajuan[7] solved the functional linear regression model by using functional principal component estimation and used Lasso to select the characteristic function to predict the return rate. Song Xiaofeng [8] optimized the model and accurately predicted the air quality based on ridge regression.

In this paper, a prediction method combining the model averaging and principal component regression is applied to time series data. The significant advantage of this method is that it considers the potential relationship between the response and principal component scores, and avoids the choice of the number of principal components.

2. Method and theory

2.1 Principal component regression

Principal component analysis [9] is known as a method of dimension reduction, multiple highly correlated variables in the original data set are transformed into a group of new independent or unrelated variables by linear combination. Suppose that there are n samples and each sample has p data, then the $n \times p$ dimensional matrix is formed:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{1}$$

The eigenvalues and eigenvectors of the matrix can be calculated by Jacobian method[10]: λ and a_{ij}. The eigenvalue λ_i corresponds to the projected variance of each principal component, which can reflect the amount of information provided by every component. When λ is larger, more information is contained. i = 1,2,⋯,p, principal component can be expressed as:

$$\begin{cases} F_1 = a_{11}Z_1 + a_{12}Z_2 + \cdots + a_{1p}Z_p \\ F_2 = a_{21}Z_1 + a_{22}Z_2 + \cdots + a_{2p}Z_p \\ \cdots \\ F_i = a_{p1}Z_1 + a_{p2}Z_2 + \cdots + a_{pp}Z_p \end{cases} \tag{2}$$

where Z_p is the Standardized predictor variables. We calculate the cumulative contribution rate and determine the number of principal components k. The k factors with the cumulative contribution rate of more than 80% were selected as explanatory variables of the multiple linear regression equation and the least square method was used to solve the parameters:

$$y = \beta_0 + \beta_1 F_1 + \beta_2 F_2 + \cdots + \beta_k F_k \tag{3}$$

Finally, transforming the solved model into the principal component regression model based on the original variables by matrix transformation.

2.2 Model averaging based on Bayes (BMA)

The principle and procedure of Bayesian model averaging are as follows: In the first step, BMA randomly combines the principal components processed by PCA. For a dataset containing k explanatory variables, they can be combined to generate K possible linear regression models, whose model space is called M. The second step is to calculate the posterior probability. On the premise of obtained data D, the prior probability of each model is set as $P(M_k)$ and the prior distribution to solve the posterior probability[11]:

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{i=1}^{K} P(D|M_I)P(M_I)} \tag{4}$$

Among them: $P(D|M_k) = \int P(D|\theta_k, M_k)P(\theta_k|M_k)d\theta_k$, $P(D|M_k)$ is the marginal likelihood function of model M_k, θ_k is the parameter of M_k, $P(\theta_k|M_k)$ is the prior density function of θ_k in model M_k, $P(\theta_k|M_k)$ is the likelihood function. Then, the weights of the models in the model class are determined by the posterior model probability and adjusted according to the degree of influence of the given variables on the results. In order to avoid overfitting of the model to the training set, Bayesian information Criterion (BIC) is used to penalize the model. BIC [12] was defined as:

$$BIC_k = kln(n) - 2ln(L) \tag{5}$$

Among them, k is the number of model parameters, n is the number of samples, and L is the maximum likelihood function. Determine the combined weights for each model:

$$w_k = \frac{exp(-BIC_k/2)}{\sum_{k=1}^{K} exp(-BIC_k/2)} \tag{6}$$

2.3 Principal component regression based on the model averaging

The selection of principal components plays a key role in the establishment of regression model. The combination of principal components can change the parameter size and prediction accuracy of the prediction model. Therefore, the idea of model averaging is combined with the principal component regression algorithm to generate an adaptive model combining different explanatory variables for prediction to avoid the risk of poor prediction stability of regression model in high-dimensional time series data. We use a relatively simple model averaging method based on Bayes to build a model and compare its advantages in high dimensional data prediction.

Firstly, using PCA to reduce the dimension of data. For fear of losing the information related to the response variables, the projection data with the variance contribution rate up to 99.99% were extracted for combined prediction. Specifically, K regression models were generated by random combinations of principal components in the projection data, and their prior probabilities and prior distributions of each parameter were determined.

Secondly, the BIC information criterion was used to punish these K models to enhance the generalization ability of the final model.

Finally, the combined weight of each model is calculated by the posterior model probability and the weighted average is obtained to get the final prediction result. Figure 2 shows the average based on the Bayesian model of principal component regression flow chart:

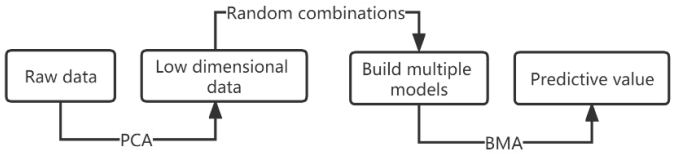

Figure 2: Principal component regression based on Bayesian model average

3. Data analysis

3.1 Data source and processing

The data comes from the website：http://lib.stat.cmu.edu/datasets/tecator where data is recorded from Tecator Infratec Food and feed analyzer. The dataset contains a total of 215 meat samples. Each meat sample consists of absorbance at 100 different wavelengths and the content of fat, water and protein. The basic experimental process is shown in Figure 3:

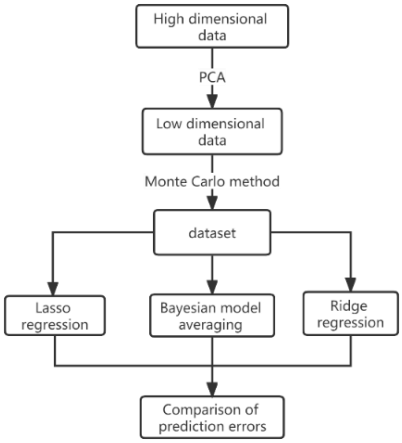

Figure 3: Basic steps for meat content prediction

3.2 Comparison experiment

Because the original data contains three reaction variables, namely water, protein and fat, three data sets were generated respectively. Firstly, the dimensionality of each dataset was reduced by PCA, and the projection data with a cumulative contribution of 99.99% was selected. After PCA processing, the predictive variable matrix of each dataset was converted from $X_{215\times100}$ to $X_{215\times6}$. In other words, six principal components of information were extracted from 100 spectral absorption rate variables after PCA treatment as explanatory variables in the next regression analysis.

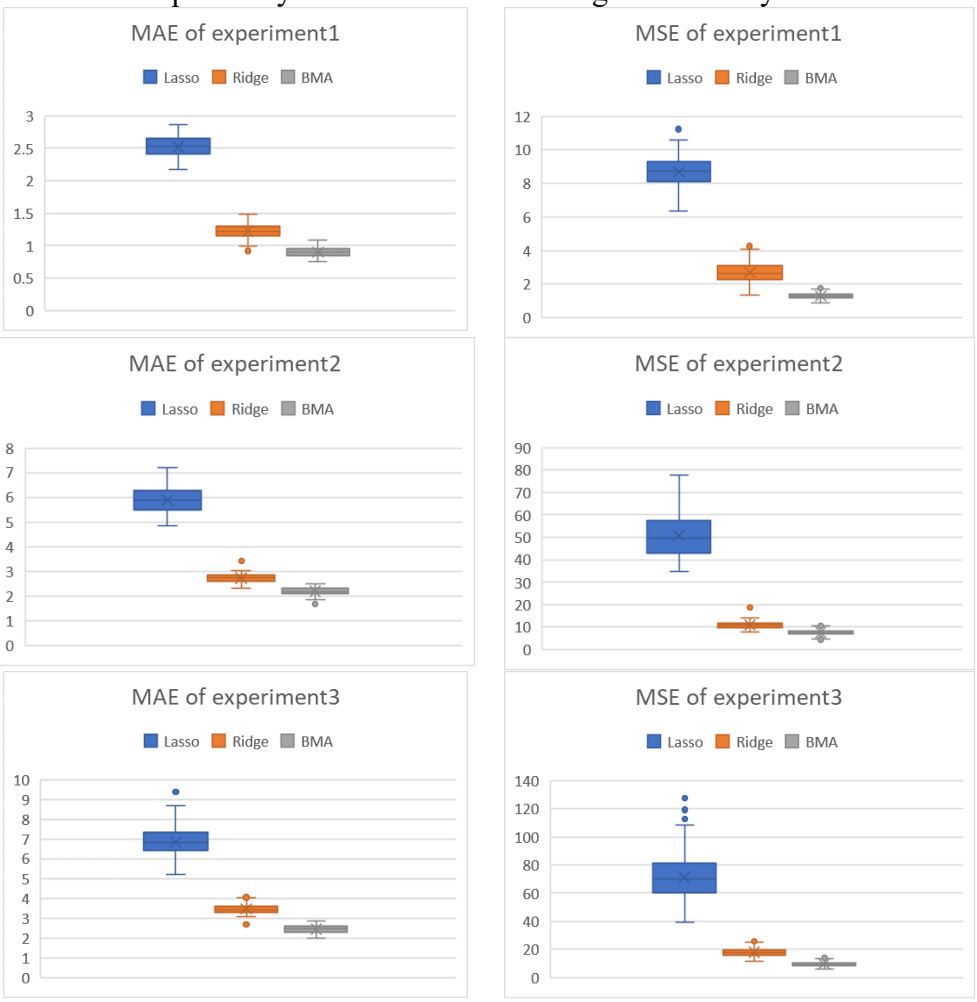

Figure 4: Prediction error of 100 simulation experiments

Secondly, the data was split into training set and test set according to the ratio of 7 to 3, and used for model fitting and testing respectively. Then, Bayesian model averaging, Lasso regression and Ridge regression analysis were performed on the three datasets by using Python. The mean absolute error (MAE) and mean square error (MSE) between the output prediction results and the actual results were calculated respectively. We use Monte Carlo algorithm to generate 100 sets of simulated data sets for each of the three data sets and repeated the regression prediction experiment 100 times to verify the stability and correctness of the model prediction. The prediction accuracy of the model reflects the prediction accuracy and generalization ability. Among the prediction errors obtained by repeated training of the three models, the average prediction error of the Bayesian model averaging method is the smallest, while the Lasso regression method is the largest. The prediction errors of 100 simulated experiments of the three task sets are shown in the box plot in Figure 4. The two types of

prediction errors of the three datasets were analyzed respectively, and it was found that the MAE and MSE values of the model averaging method were less than 10, and their variances were within the interval (0,2), indicating that the model has high accuracy and low risk in the prediction process. In comparison, this algorithm is more stable. Table 1-3 display the prediction error results for meat water, fat and protein content prediction:

The experimental results show that the prediction error of principal component regression based on Bayesian model average is significantly smaller than that of Lasso regression and ridge regression, which has the highest prediction accuracy and the strongest generalization ability. Good prediction results were obtained in three different meat task sets. Therefore, this algorithm is superior to Lasso regression and Ridge regression when predicting high-dimensional time series data.

Table 1: Experiment-protein

	mean of MAE	mean of MSE	std of MAE	std of MSE
BMA	0.8997042	1.278986	0.076394	0.1905179
Lasso	2.527961	8.692473	0.144164	0.933198
Ridge	1.224671	2.689623	0.108893	0.572077

Table 2: Experiment-water

	mean of MAE	mean of MSE	std of MAE	std of MSE
BMA	2.185745	7.542573	0.15053	1.275683
Lasso	5.891213	50.843238	0.524787	9.384385
Ridge	2.710122	10.877201	0.175696	1.574278

Table 3: Experiment-fat

	mean of MAE	mean of MSE	std of MAE	std of MSE
BMA	2.449645	9.694131	0.1914911	1.856409
Lasso	6.836761	7-.984008	0.779037	17.200226
Ridge	3.460719	18.054175	0.270012	3.116801

4. Conclusion

Experiments found that in dealing with high dimension and less sampled data, using principal component analysis to under the condition of the least loss of information to convert data into low dimension space is analyzed. In the principal component regression forecasting model, the number of principal components is more reliant on the cumulative contribution rate for selection. It is easy to cause the previous data to overwrite the amount of information behind. In addition, the establishment of the model does not take into account the uncertain relationship between principal components and response variables, so it is risky to use the model established by principal component regression to make predictions. We only use the model averaging method based on Bayes criterion to prove that this algorithm can predict without relying on the selected model and fully consider the uncertain relationship between principal components and response variables. Therefore, the results of time series prediction using the information criterion with more strict punishment on the model could be more accurate. The algorithm can be applied to high-dimensional time series data in different scenarios to adaptively train a model with high prediction accuracy.

References

[1] Zhao Xun. Principal component analysis and neural network application in consumer spending forecast [D]. Jilin University, 2016.

[2] Zhang He, Fan Mengxuan. Based on Lasso regression model analysis of Qingdao Marine economy and Marine industry [J]. Journal of ocean development and management, 2022, 33 (8): 6. 22 to 28 DOI: 10.20016/j.carol carroll nki hykfygl. 20220803.002.
[3] Shi Y. Research on soil composition prediction model based on visible near-infrared spectroscopy [D]. University of Science and Technology of China, 2018.
[4] Meng Qinglong, Shang Jing, Huang Renshuai, Zhang Yan. Prediction model of apple soluble solid content based on principal component regression [J]. Preservation and Processing, 2020, 20(05): 185-189.
[5] Zhu Hailong, Li Pingping. Analysis of Financial Revenue Influencing factors in Anhui Province based on Ridge regression and LASSO regression [J]. Lancet journal of jiangxi university of science and technology, 2022 (01): 59-65. DOI: 10.13265/j.carol carroll nki JXLGDXXB. 2022.01.009.
[6] Xu Yunjuan, Luo Youxi. The Lasso dimension reduction based on variable clustering algorithm and simulation [J]. Journal of statistics and decision, 2021, 5 (4): 31-36 DOI: 10.13546/j.carol carroll nki tjyjc. 2021.04.007.
[7] Li Y J. Research on functional principal component regression model and yield prediction based on LASSO [D]. Xiamen University, 2019.
[8] Song Xiaofeng. Air quality index based on ridge regression prediction [J]. Journal of electronic world, 2020 (15): 87-88. The DOI: 10.19353 / j.carol carroll nki DZSJ. 2020.15.046.
[9] Yuan Yuliang, Sheng Wenyi. Prediction method of stem diameter dynamic change based on principal component regression [J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(01): 306-314.
[10] Sun Jianhua, Zhang Zhili, Shi Qian, Zhao Yang, Wei Chunrong. Study on prediction of gas emission based on principal component stepwise regression analysis [J]. Coal Engineering, 2020, 52(01): 89-94.
[11] Xie Changye. Research on Monte Carlo Option Pricing based on Bayesian Model Averaging (BMA) method [D]. Nanjing University,2018.
[12] Zhang Xinyu, Zou Guohua. Model average method and its application in prediction [J]. Journal of statistical research, 2011, 28 (6): 97-102. The DOI: 10.19343/j.carol carroll nki.11-1302/c. 2011.06.018.