ISSN 2616-5767 Vol.6, Issue 5: 23-29, DOI: 10.25236/AJETS.2023.060505

A Time Series Regression Model via Improved PCA and Bagging Algorithms

Jinzhi Fan[1], Ziyi Fan[2]

[1]*Changsha University of Science & Technology, Changsha 410114, China*
[2]*Changzhi University, Changzhi, 046011, China*

Abstract: *Considering the case that the prediction variable is a time series and the response variable is a continuous scalar, we propose a time series regression model based on improved PCA and Bagging Algorithms. Compared with PCA dimension reduction, the proposed method uses distance correlation coefficient matrix instead of Person correlation coefficient matrix, which makes the distribution assumption of original variables more free. Considering that PCA is an unsupervised dimension reduction technique and the connection functions between principal components and response variables are unknown, we propose to use Bagging Algorithmss to capture information of principal components related to response variables. In the actual data analysis, the comparative methods are LASSO and PCA-based linear models, and the empirical results show that the proposed method has certain competitiveness compared with the comparison method.Finally, because the base-model of Bagging Algorithms is model-free, some machine learning methods with higher precision and flexibility can be used as the base-model for data tasks with different complexity.*

Keywords: *Time Series Data Regression, Distance correlation coefficient, PCA, Bagging Algorithms*

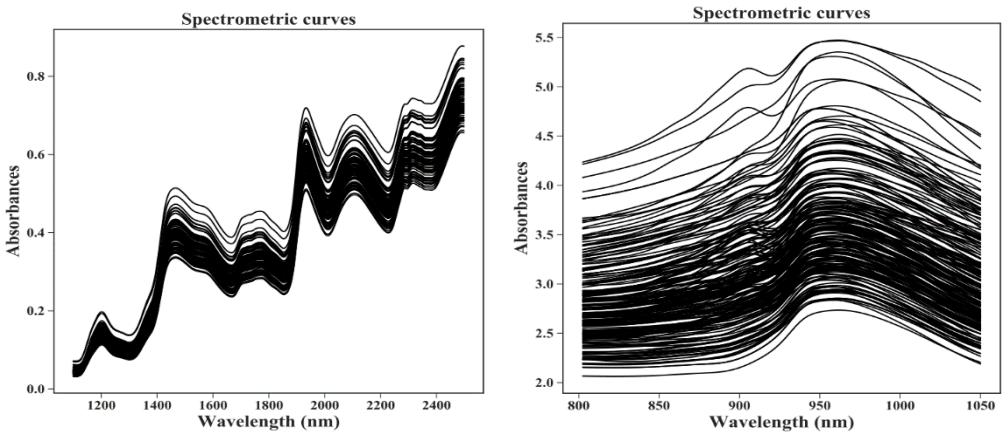
1. Introduction

Time series (or dynamic series) refers to a series that arranges the values of the same statistical index according to the time sequence of its occurrence, which can be divided into three categories: absolute logarithmic time series, relative number time series and mean time series. The problem of regression with time series as the predictive variable is called time series regression problem.We use two examples to illustrate the data background. First of all, in the meat dataset, the prediction variable is the absorption spectrum curve measured by the near-infrared spectroscopy analyzer, which is expressed as a time series, as shown in Figure 1-a, the response variables are the moisture, fat and protein content of meat, and the task is to predict the moisture, fat and protein content of meat according to the spectral curve; Secondly, in the maize dataset, the prediction variables is the absorption spectrum data measured by the near-infrared spectrometer, see Figure 1-b, and the response variables are the oil, moisture and starch content of maize. The task is to predict the oil, moisture, and starch content of corn based on the spectral curve.

Multiple linear regression is a mathematical model for regression analysis, which is widely used in economy, agriculture and so on. Haowen Dong et al. [1] used multiple linear regression method to make regression prediction of wind speed with five factors such as air pressure and wind Angle as response variables. Xuejian Li et al. [2] used multiple stepwise regression analysis to study 12 major factors affecting soil moisture and determined a drought detection model. Ji Shu et al.[3] studied rice leaf area index and established an estimation model based on feature belt and vegetation index through ridge regression and multiple stepwise regression. However, multiple linear regression models have dimensionality disaster problems when the dimensionality becomes higher [4]. Shukui Bo et al. [5] studied related machine learning problems that deteriorate dramatically with the increase of dimensionality. In multiple regression models, principal component analysis (PCA) is often used for dimensionality reduction. Xiangwei Kong et al. [6] used PCA to reduce the dimensionality of seven indexes such as lithological density in the drill bit optimization problem, and obtained the most representative parameters to characterize the drillability extremes, and realized the prediction of the drillability extremes of hard formation rocks. Yunfei Zhang et al.[7] based on the PCA-SSA-ELMAN prediction model, the dimensionality reduction treatment of meteorological factors and the AQI value of the air quality index were predicted. Bingchen He et al. [8] analyzed and reduced dimensions of multiple factors affecting battery life, such as capacity, and established a PCA-GPR model to predict the remaining service life of lithium-ion batteries. However, PCA is a dimensionality reduction method based on Pearson correlation

ISSN 2616-5767 Vol.6, Issue 5: 23-29, DOI: 10.25236/AJETS.2023.060505

coefficient. The correlation coefficient of two variables is zero, which means that they are not linearly correlated, but it is impossible to determine whether there is a nonlinear correlation. In addition, PCA is an unsupervised dimensionality reduction technique, and the link function between each principal component and the response variable is unknown, which greatly reduces the efficiency of completing the prediction task. In view of the two problems existing in the original PCA, firstly, we propose to use the distance correlation coefficient matrix instead of the covariance matrix, which has the advantage of replacing the Pearson correlation coefficient with the distance correlation coefficient, which reduces the constraints on the original variables and makes the distribution assumptions of the original variables more relaxed. Secondly, in order to excavate the connection structure between principal components and response variables, Bagging model based on Bootstrap method is used to capture the effective relationship between them, and the deviation and variance of the prediction model is effectively balanced by multi-model fusion [9]. The actual data analysis shows that the proposed method is more efficient than the comparison method.



(a) Spectral curve of meat *(b) Spectral curve of corn*

Figure 1: Spectral plot

2. Theory and Method

2.1 PCA based on distance correlation coefficient

In 1933, Hotelling proposed principal component analysis[10], which is a multivariate statistical analysis method that uses a dimensionality reduction idea to transform multiple indicators into several comprehensive indicators by orthogonal rotation under the premise of losing little information, and each principal component is a linear combination of the original variables, and the principal components are not related to each other.Let the p-dimensional random variable as $X=(X_1,X_2,...,X_p)$,the mean of the random variable X as μ and the covariance matrix as Σ.Principal component analysis obtains p principal components $Y_1,Y_2,...,Y_p$ by performing linear transformations.

$$\begin{cases} Y_1=\mu_{11}X_1+\mu_{12}X_2+...+\mu_{1p}X_p \\ Y_2=\mu_{21}X_1+\mu_{22}X_2+...+\mu_{2p}X_p \\ ..... \\ Y_p=\mu_{n1}X_1+\mu_{n2}X_2+...+\mu_{np}X_p \end{cases} \tag{1}$$

The specific process is as follows:

Step 1: Standardize raw data;

Step 2: Calculate a matrix of correlation coefficients between standardized variables;

Step 3: alculate the eigenvalues and eigenvectors of the correlation coefficient matrix;

Step 4: Calculate the contribution rate of each principal component;

ISSN 2616-5767 Vol.6, Issue 5: 23-29, DOI: 10.25236/AJETS.2023.060505

Step 5: Extract principal components based on cumulative contribution rates.

Szekely et al. proposed a method for measuring the correlation between distance covariance and distance correlation coefficient [11], which is abbreviated as $dCov(X,Y)$ and $dCor(X,Y)$, respectively. Where $dCor(X,Y)=0$ indicates that X and Y are independent of each other. In contrast to the Pearson correlation coefficient, the distance correlation coefficient mainly measures the degree of nonlinear correlation. For two n-dimensional variables, the distance correlation coefficient is calculated as follows:

$$dCor(X,Y)=\frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}} \tag{2}$$

Where,

$$dCov^2(X,Y)=\frac{1}{n^2}\sum_{j,k=1}^{n}(a_{j,k}-\bar{a}_{j.}-\bar{a}_{.k}+\bar{a}_{..})(b_{j,k}-\bar{b}_{j.}-\bar{b}_{.k}+\bar{b}_{..}) \tag{3}$$

$$a_{j,k}=\left\|X_j-X_k\right\| \quad j,k=1,2,...,n \tag{4}$$

$$b_{j,k}=\left\|Y_j-Y_k\right\| \quad j,k=1,2,...,n \tag{5}$$

$$dVar^2(X)=dCov^2(X,X) \tag{6}$$

$$dVar^2(Y)=dCov^2(Y,Y) \tag{7}$$

When Pearson correlation coefficient in PCA analysis changes to distance correlation coefficient and other analysis methods do not change, principal component analysis based on distance correlation coefficient is generated, that is, covariance in PCA changes to distance covariance.

2.2 Bagging Algorithmss

The bagging Algorithms is a technique that reduces generalization errors by combining multiple models. The main idea is self-sampling and voting[12].The flow chart is shown in Figure 2 and the process is as follows:

Step 1: Extract training set from original sample set. In each round, n training samples were extracted from the original sample set using the Bootstraping method. A total of k rounds were extracted to obtain k training sets.

Step 2: Use one training set each time to get a model, k training sets to get a total of k models. We can use different classification or regression methods according to specific problems, such as decision tree, perceptron, etc.

Step 3: For the classification problem: the k models obtained in the previous step are voted to get the classification results; For the regression problem, the mean value of the above model is calculated as the final result.

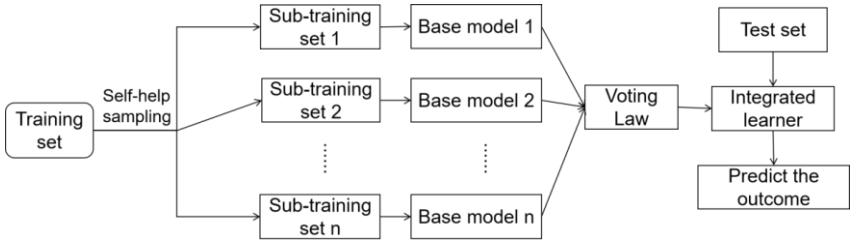


Figure 2: The process of Bagging

ISSN 2616-5767 Vol.6, Issue 5: 23-29, DOI: 10.25236/AJETS.2023.060505

2.3 Time series regression based on improved PCA and Bagging Algorithms based on distance coefficient matrix

In the previous chapters, we considered introducing the distance correlation coefficient matrix into PCA to build an improved PCA based on the distance coefficient matrix. A subsequent problem is how to construct effective connection functions between each principal component and response variables. Here, since the underlying model structure is unknown, we use the idea of model average processing. Specifically, we will use Bagging Algorithms to construct multiple sub-models between the principal component and response variables, so that the deviation and variance of the prediction model can be balanced. The specific model process is shown as follows:

Step 1: Divide the data into training set and test set; Defined as A and B;

Step 2: The improved PCA was used to calculate the projection matrix on the training set A, and the training set data C and test set data D were obtained after dimensionality reduction;

Step 3: The data set C is brought into the Bagging Algorithms for training to obtain the parameters of each sub-model. Finally, the data set D is brought into the trained Bagging Algorithms to obtain the prediction effect on the training set, denoted as Y1;

Step 4: The accuracy of the Algorithms is calculated using the evaluation index and compared with the comparison method.

3. Data Analysis

3.1 Data sources and descriptions

To test the effectiveness of the method used in this paper, two data sets of meat and corn are selected. Meat data set from **http://lib.stat.cmu.edu/datasets/tecator/**, a total of 204 samples,the prediction variables are 100 absorption spectrum data measured by the near-infrared spectroscopy analyzer in the band of 850~1050 nm, and the response variables are the moisture, fat and protein content in meat.The maize dataset is derived from **https://eigenvector.com/data/Corn/**, with a total of 80 samples, the prediction variables are 700 absorption spectrum data measured by the near-infrared spectroscopy analyzer in the band of 1100~2498 nm, and the response variables are the oil, moisture and starch content in the maize.

3.2 Empirical results

We use the improved PCA based on distance correlation coefficients and Bagging Algorithms to regress the meat and corn time series datasets, the first comparison method is LASSO, and the second comparison method is linear regression after dimensionality reduction using PCA.The effects of the three methods are analyzed by using three types of errors: mean squared error MSE, absolute error MRE and posterior error BE, and the calculation formula of the three types of error is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{8}$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{9}$$

$$BE = \frac{std(y-\hat{y})}{std(y)} \tag{10}$$

The implementation environment is Python. Firstly, 100 simulation experiments are carried out on meat data to obtain the MSE, MRE and BE corresponding to the three methods, and then the mean and standard deviation of various errors are obtained. The error data for fat and protein content in meat are shown in Tables 1, 2 and 3, respectively.

ISSN 2616-5767 Vol.6, Issue 5: 23-29, DOI: 10.25236/AJETS.2023.060505

Table 1: Incorrect data on the amount of fat in meat

index method	DPCA		LASSO		PCA	
	Mean value	Standard deviation	Mean value	Standard deviation	Mean value	Standard deviation
MSE	3.71	0.28	4.20	0.32	13.99	5.68
MRE	3.05	0.20	3.43	0.24	10.95	4.40
BE	0.19	0.02	0.33	0.03	1.10	0.45

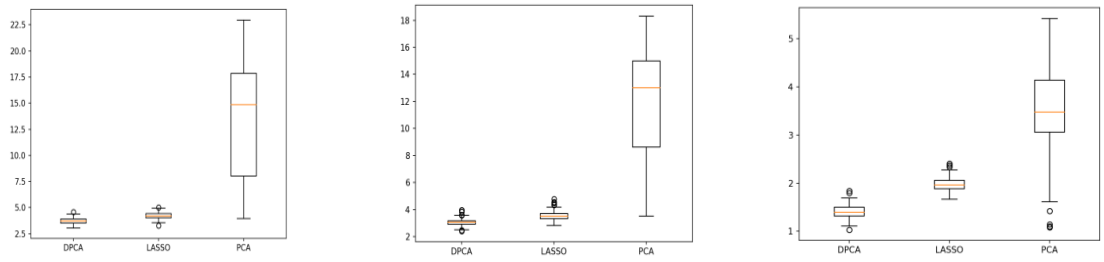
Table 2: Incorrect data on the amount of water in meat

index method	DPCA		LASSO		PCA	
	Mean value	Standard deviation	Mean value	Standard deviation	Mean value	Standard deviation
MSE	3.07	0.26	3.55	0.35	11.36	4.52
MRE	2.50	0.19	2.91	0.24	8.97	3.55
BE	0.31	0.03	0.36	0.03	1.14	0.46

Table 3: Incorrect data on the amount of protein in meat

index method	DPCA		LASSO		PCA	
	Mean value	Standard deviation	Mean value	Standard deviation	Mean value	Standard deviation
MSE	3.71	0.28	4.20	0.32	13.99	5.68
MRE	3.05	0.20	3.43	0.24	10.95	4.40
BE	0.29	0.02	0.33	0.03	1.10	0.45

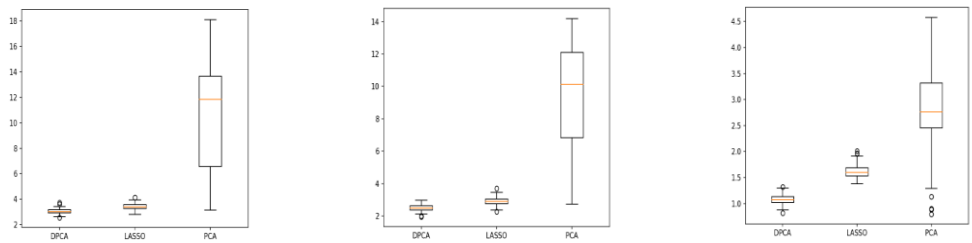
For the fat content in meat, the mean square error MSE of the improved PCA and Bagging Algorithms is 0.49 and 10.28 lower than that of the comparison methods 1 and 2, respectively.The mean value of absolute error MRA is 0.38 and 7.9, respectively. The mean value of the posterior error BE is 0.14 and 0.91, respectively.For water content and protein content, improved PCA and Bagging Algorithms methods are more effective. We visualized the data and got the error box plot. The box plots of MSE corresponding to the fat, water and protein contents in meat are shown in Figure 3-a, 3-b and 3-c respectively.



(a) MSE of fat *(b) MSE of water* *(c)MSE of proteins*

Figure 3: Boxplot of MS

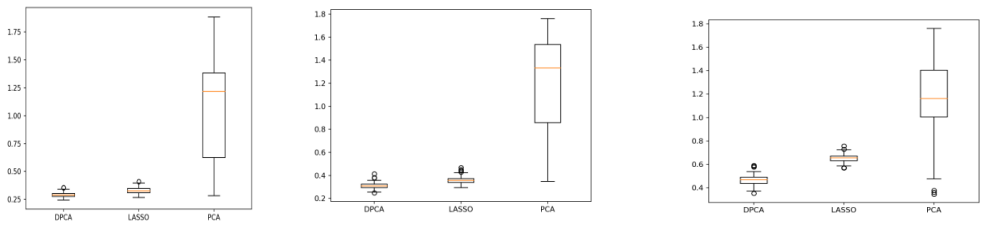
See Figure 4-a, 4-b and 4-c for the MRE box plots corresponding to the fat, water and protein contents in meat. See Figure 5-a, 5-b, and 5-c for the corresponding BE box plots of fat, water, and protein contents in meat.



(a) MRE of fat *(b) MSE of water* *(c) MSE of proteins*

Figure 4: Boxplot of MRE

ISSN 2616-5767 Vol.6, Issue 5: 23-29, DOI: 10.25236/AJETS.2023.060505



(a) BE of fat *(b) BE of water* *(c) BE of proteins*

Figure 5: Boxplot of BE

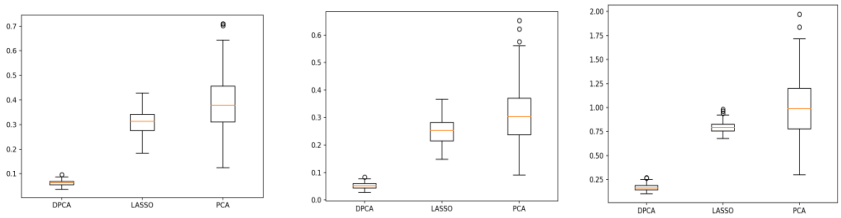
For corn data, first of all, 100 simulation experiments are conducted, response variables are used to regression oil content, and MSE, MRE and BE corresponding to the three methods are obtained, and then the mean and standard deviation of various errors are obtained. The error data for the oil content in corn is shown in Table 4.

Table 4: Incorrect data on oil content in corn

index method	DPCA		LASSO		PCA	
	Mean value	Standard deviation	Mean value	Standard deviation	Mean value	Standard deviation
MSE	0.06	0.01	0.31	0.05	0.39	0.12
MRE	0.05	0.01	0.25	0.05	0.31	0.11
BE	0.16	0.03	0.79	0.06	1.01	0.34

For the oil content in corn, the mean square error MSE of the improved PCA and Bagging Algorithms is 0.25 and 0.33 lower than that of the comparison methods 1 and 2, respectively. The mean absolute error of MRE is 0.2 and 0.26, respectively. The mean value of the posterior error BE is 0.63 and 0.85, respectively. Similarly, we visualize the data and get the boxplot of various errors. The box plots of MSE, MRE and BE corresponding to corn oil content are shown in Figure 6-a, 6-b and 6-c respectively.

The proposed method shows better results in the above four regressions, which may be because the original data does not obey normal analysis. PCA based on distance correlation coefficient is used for dimensionality reduction to obtain mutually independent principal components, which improves the prediction accuracy. It may also be because the Bagging Algorithms is used to capture the information related to the principal component and response variables, which has the advantages of model average and model freedom.



(a) MSE of oil *(b) MRE of oil* *(c) BE of oil*

Figure 6: Box plot of each error of oil content

4. Conclusion and Prospect

Considering that the prediction variable is a time series and the response variable is a continuous scalar, we propose a time series regression model based on improved PCA and Bagging Algorithms. The proposed method uses distance correlation coefficient instead of Pearson correlation coefficient to reduce the constraint on the original variable distribution. In addition, based on the idea of model average, the deviation and variance of the prediction model are effectively balanced through the fusion of multiple models.

As for the future research direction, firstly, considering that Bagging Algorithms is a model averaging method, that is, all models have the same weight, but different weights of all models may be required in practical problems, relevant weighting criteria of model average can be introduced into Bagging, so as to

ISSN 2616-5767 Vol.6, Issue 5: 23-29, DOI: 10.25236/AJETS.2023.060505

generate weighted average Bagging Algorithms and improve Algorithms efficiency. For example, autonomous model averaging, optimized model averaging and heuristic model averaging; Secondly, since the base regression method of Bagging Algorithms is model free, machine learning methods with higher accuracy and flexibility can be used as the base model when dealing with different complex data, such as neural networks. Finally, because PCA selects the final principal component based on the cumulative contribution rate of each principal component, some principal components with small contribution rate but large impact on the response variables are not selected, new evaluation indexes can be introduced on the basis of the cumulative contribution rate, and principal components can be selected by combining the cumulative contribution rate and influence.

References

[1] Haowen Dong, Qianying Zhang, Zhaoyi Chen, Yi Zhao, Yuqing Qian. Comprehensive Prediction of Wind Speed Capture in Wind Power Generation Based on Multiple linear regression and time Series [J]. Mechanical and Electrical Information, 2020, No.633(27): 8-9. 2020.27.004.

[2] Xuejian Li, Meili Wang, Min Zhao, Yinghan Shi, Qiang Gao. Vineyard drought monitoring model based on multiple stepwise regression analysis [J]. Agricultural Research in the Arid Areas, 2022, 40(04): 249-254.

[3] Ji Shu, Gu Chen, Xi Xiaobo, Zhang Zhenghua, Hong Qingqing, Huo Zhongyang, Zhao Haitao, Zhang Ruihong, Li Bin, Tan Changwei. Quantitative Monitoring of Leaf Area Index in Rice Based on Hyperspectral Feature Bands and Ridge Regression Algorithms [J]. Remote Sensing, 2022, 14(12).

[4] Wang Y, Xia S T. A novel feature subspace selection method in random forests for high dimensional data[C]// 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016.

[5] Ravi K K V, Agrawal D, Abbadi A E. Dimensionality Reduction for Similarity Searching in Dynamic Databases [J]. Computer Vision and Image Understanding, 1999, 75(1/2): 59-72

[6] Xiangwei Kong, Hao Chen, JiajieYe, Yadong Li, Gan Yifeng. Drill bit optimization for predicting drillability of rock based on PCA [J]. Xinjiang Oil and Gas, 2022, 18(03): 6-11.

[7] Yunfei Zhang, Wanxiong Wang. Prediction of Air Quality Index in Xi'an Based on PCA-SSA-Elman [J]. Journal of Software, 2002, 43(06): 30-34.

[8] Bingchen He, Xueming Yang, Jinsong Wang, Xu Zhu, Hu Zongjie, Liu Qiang. Prediction of Residual Service Life of Lithium ion Batteries based on PCA-GPR [J]. Acta Energiae Solaris Sinica, 2022, 43(05): 484-491. 2022-0422.

[9] Rong Zhu, Guohua Zou, Xinyu Zhang. Model averaging Method for Partial Function Linear Models [J]. Journal of Systems Science and Mathematics, 2018, 38(07): 777-800.

[10] Xiaoqun He. Applied Regression Analysis (R Language Edition) [M]. Beijing: Publishing House of Electronics Industry, 2017.

[11] Lu Zhang, Lingchen Kong, Huangyue Chen. Hierarchical clustering method based on Distance Correlation Coefficient [J]. Computational Mathematics, 2019, 41(03): 320-334.

[12] Xie Q, Tang L, Li W , et al. Principal Model Analysis Based on Partial Least Squares[J]. 2019.