Time series forecasting of stock market index based on ENW-MPCA model

Zhiqiang Guo\textsuperscript{1,2}\textsuperscript{*}, Xiao Wang\textsuperscript{1,2}, Yali Zeng\textsuperscript{1,2} and Jie Yang\textsuperscript{1,2}

\textsuperscript{1}School of Information Engineering, Wuhan University of Technology, Wuhan, China
\textsuperscript{2}Hubei Key laboratory of Broadband Wireless Communication and Sensor Networks, Wuhan University of Technology, Wuhan, China
\textsuperscript{*}Corresponding author’s e-mail: guozhiqiang@whut.edu.cn

Abstract. Aiming at the high dimension of factors affecting the stock market index, and the problem that MPCA algorithm cannot distinguish the contribution of different eigenvectors to financial time series forecasting, eigenvalue normalization weighted multilinear principal component analysis (ENW-MPCA) is put forward. Due to the existence of correlation between stock markets, 34 technical indexes of 7 stock markets are selected to construct a three-dimensional tensor model. Eigenvectors of different eigenvalues have different contributions to the prediction accuracy of MPCA algorithm, thus adopt ENW-MPCA algorithm for the feature extraction, which can distinguish the different contribution of eigenvectors corresponding to different eigenvalues to the prediction while reducing dimension. Then using support vector machine for regression prediction, the prediction value of financial time series is obtained. Experimental results on the Hang Seng Index shows that, compared with the MPCA, the forecasting error of ENW-MPCA is smaller and the prediction accuracy is improved to a certain extent. This indicates that the proposed algorithm can fully retain the internal structure of stock time series, proving its validity and practicability.

1. Introduction

As main body of financial market, stock market plays an important role in development of national economy. The change of the stock market index reflects the law of the stock market to a certain extent. Therefore, time series forecasting of stock market index has become a hot spot for both investors and academic researchers\cite{1-4}.

Due to the characteristics of high dimension and numerous related indexes of stock market index time series, directly using the prediction model for classification or regression often lead to the dimension disaster, greatly increasing the computational complexity. Therefore, many existing methods combine feature extraction algorithm and prediction model to predict\cite{5}. Grigoryan.H et al.\cite{6} put forward an integrated forecasting model (ICA-SVM) based on ICA and SVM. Guo.Z et al.\cite{7} proposed a time series prediction method of stock market index based on feature fusion, combining independent component analysis and canonical correlation analysis to extract and fuse features, and then SVM model was adopted to predict. Using the correlation between related stock markets, after building a three-dimensional tensor model with a certain amount of input indexes, Guo.Z et al.\cite{8} extracted features of the model by using multilinear principal component analysis (MPCA). And then regression prediction was carried out, and good results were obtained.
However, MPCA algorithm treats the eigenvectors of each dimension equally, and does not distinguish the influence of each eigenvector in the feature matrix on the feature extraction. Therefore, in view of the fact that the MPCA algorithm cannot distinguish the influences of different eigenvectors on the prediction of financial time series, this paper proposes a eigenvalue normalization weighted multilinear principal component analysis (ENW-MPCA). The contribution of eigenvectors of different eigenvalues to the prediction accuracy is distinguished by weighting the eigenvectors of the projection matrix in all the modes of the tensor. It can not only preserve the characteristics of the original MPCA algorithm, but also increase the prediction accuracy of the algorithm by adding feature weights.

2. Research methodology

2.1. Multilinear principal component analysis (MPCA)

MPCA is a multilinear method for dimensionality reduction in all tensor modes, seeking bases in each mode that allow projected tensors to capture the majority of variation of the original tensor. It captures most of the changes of the original tensor, including the structural information and correlation, so it can achieve dimension reduction while avoiding the destruction of correlation between the sample data of PCA and overcome the shortcoming of the destruction of data structure.

For tensor sample $\mathbf{X}_m \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, where $N$ is the order of tensor and $m = 1, \ldots, M$, of which $M$ is number of samples, the centralization is $\mathbf{X}_m = \mathbf{X}_n - \bar{\mathbf{X}}$, where $\bar{\mathbf{X}} = \frac{1}{M} \sum_i \mathbf{X}_i$. $n$-mode unfolding of the tensor $\mathbf{X}^{(n)}$ is denoted as $\mathbf{X}^{(n)}_{m(n)} = \mathbb{R}^{I_1 \times (I_n - I_{n-1}) \times I_{n-1}}$ and the scatter matrix of each mode can be measured by equation (1):

$$\psi^{(n)} = \sum_{m=1}^{M} \mathbf{X}^{(n)}_{m(n)} \mathbf{X}^{(n)}_{m(n)\mathsf{T}} (n = 1, \ldots, N)$$

$$U^{(n)} \in \mathbb{R}^{I_n \times P_n} (P_n < I_n)$$ can be measured by computing the eigenvectors of the scatter matrix analytically.

Objective function of MPCA, which is to determine $N$ projection matrices $\{U^{(n)}, n = 1, \ldots, N\}$ that maximizes total dispersion $\psi_y$ by iteration, satisfy equation (2):

$$U^{(n)} = \arg \max \psi_y$$

By using linear transformation $\{U^{(n)}, n = 1, \ldots, N\}$, project the tensor to subspace $\mathbb{R}^I \otimes \mathbb{R}^P \otimes \cdots \otimes \mathbb{R}^P$ to obtain $\mathbf{y} \in \mathbb{R}^{P_1 \times P_2 \cdots P_N}$ so as to achieve dimension reduction:

$$\mathbf{y} = \tilde{\mathbf{X}} \times_1 U^{(1)\mathsf{T}} \times_2 U^{(2)\mathsf{T}} \cdots \times_N U^{(N)\mathsf{T}}$$

2.2. ENW-MPCA algorithm

Aiming at the problem that MPCA algorithm cannot distinguish the contributions of different eigenvectors on the prediction of financial time series, a eigenvalue normalization weighted multilinear principal component analysis (ENW-MPCA) is proposed in this paper. By weighting the eigenvectors of projection matrix in all the modes of the tensor, the contributions of eigenvectors of different eigenvalues to the prediction accuracy are distinguished. It can not only preserve the characteristics of the original MPCA algorithm, but also increase the prediction accuracy by adding feature weights.

In this paper, the eigenvalues after normalizing are used as the weights of corresponding eigenvectors, so as to highlight the important role of eigenvectors of large eigenvalues and also reflect the auxiliary function of eigenvectors of small eigenvalues. After normalizing the eigenvalues selected
in each tensor model and diagonalizing the normalized eigenvalues, eigenvectors corresponding to the eigenvalues are multiplied by the diagonal matrices to achieve weighting.

The eigenvalue normalization formula is:

\[
\lambda_{ni}^* = \frac{\lambda_{ni}}{\sum_{n=1}^{N} \lambda_{n}} , \quad i = 1, 2, \cdots, P_{n} , \quad n = 1, 2, \cdots, N
\]

(4)

Where, \(P_{n}\) is the dimension of eigenvectors and \(N\) is the order of the tensor sample. \(\lambda_{ni}\) represents the \(i\)-th eigenvalues in the \(n\) order of the tensor sample and \(\sum_{n=1}^{N} \lambda_{n}\) represents the sum of eigenvalues chosen in the direction of \(n\) order.

For feature matrix \(U^{(n)}(n=1,\ldots,N)\), the weights are diagonal matrices consisting of \(\{\lambda_{n1}^*, \lambda_{n2}^*, \cdots, \lambda_{nP_{n}}^*\}(n=1,\ldots,N)\), which is calculated as equation (5):

\[
W_{(n)} = \text{diag} \left( \lambda_{n1}^*, \lambda_{n2}^*, \cdots, \lambda_{nP_{n}}^* \right)
\]

(5)

Therefore, the linear transformation formula of the ENW-MPCA algorithm can be reformulated by equation (6):

\[
y_{m} = \chi_{m} \times_{1} \left( W_{(1)} \times U^{(1)} \right)^{T} \times_{2} \left( W_{(2)} \times U^{(2)} \right)^{T} \cdots \times_{N} \left( W_{(N)} \times U^{(N)} \right)^{T} , \quad m = 1, 2, \cdots, M
\]

(6)

2.3. Steps of ENW-MPCA algorithm

The steps to implement the algorithm are as follows:

1. Input the tensor samples: \(\{\chi_{m} \in R^{h_{1} \times d_{2} \times d_{3}} , m = 1, \cdots, M\}\);

2. Center the samples and obtain \(\{\overline{\chi}_{m} = \chi_{m} - \overline{\chi}, m = 1, \cdots, M\}\) and \(\overline{\chi} = 1/M \times \sum_{m=1}^{M} \chi_{m}\) is the average of the sample;

3. Construct covariance matrix \(\Phi^{(n)} = \sum_{n=1}^{M} \overline{\chi}_{m(n)} \overline{\chi}_{m(n)^{T}}\) of each mode and conduct eigendecomposition, and then choose eigenvectors of the first \(P_{n}\) sorted eigenvalues to initialize the projection matrix \(U^{(n)}\), where, \(n=1,\ldots,N\);

4. Calculate the data samples after projection:

\[
\{y_{m} = \overline{\chi}_{m} \times_{1} \left( U^{(1)} \right)^{T} \times_{2} \left( U^{(2)} \right)^{T} \cdots \times_{N} \left( U^{(N)} \right)^{T} , m = 1, \cdots, M\}
\]

and \(\psi_{y} = \sum_{m=1}^{M} \|y_{m}\|_{F}^{2}\);

5. Utilize circulation to update projection matrix \(U^{(n)}(n=1,\ldots,N)\) constantly, and narrow the \(\psi_{y}\) steadily until it is less than the threshold, and output the optimal projection matrix \(U^{(n)}(n=1,\ldots,N)\);

6. Normalize the eigenvalues of eigenvectors chosen as the projection matrix:

\[
\lambda_{ni}^* = \frac{\lambda_{ni}}{\sum_{n=1}^{N} \lambda_{n}} , \quad i = 1, 2, \cdots, P_{n} , \quad n = 1, 2, \cdots, N
\]

7. Diagonalize the eigenvalues after normalization \(\hat{\lambda}_{n1}^{*}, \hat{\lambda}_{n2}^{*}, \cdots, \hat{\lambda}_{nP_{n}}^{*}\) to acquire the weights \(W_{(n)} = \text{diag} \left( \hat{\lambda}_{n1}^{*}, \hat{\lambda}_{n2}^{*}, \cdots, \hat{\lambda}_{nP_{n}}^{*} \right)\);

8. Output the expression of the samples in low-dimensional subspace:

\[
y_{m} = \chi_{m} \times_{1} \left( W_{(1)} \times U^{(1)} \right)^{T} \times_{2} \left( W_{(2)} \times U^{(2)} \right)^{T} \cdots \times_{N} \left( W_{(N)} \times U^{(N)} \right)^{T} , \quad m = 1, 2, \cdots, M
\]

3. Experimental results and analysis

In order to verify the feasibility and robustness of feature extraction of three-order tensor financial data by ENW-MPCA algorithm, experimental researches and analyses were carried out on data of the Hang Seng Index and the Nikkei 225 index, and the prediction effects were compared with MPCA algorithm and feature extraction algorithms based on vector, KPCA and PCA. The SVR of SVM
toolbox in MATLAB was adopted for regression calculation. All experiments were performed on a PC with Intel (R) Core (TM) i3-2330 CPU, 2.20G RAM memory, in MATLAB R2011b environment.

3.1. Data set preparation
The stock data collected from January 4, 2010 to December 9, 2014 contains of 34 stock indexes of 7 stock markets. Three-order tensor model is constructed and is segmented by moving sliding window. Take the empirical value 20 days for sliding window length. So the data are divided into 1000 groups and the first 800 groups are training data, and the latter are testing data. To measure the performance of the ENW-MPCA, 5 performance indicators are selected. The descriptions and formulae of these indicators are described in Table 1. Among them, y(t) is the predicted value, \( \hat{y}(t) \) is the real value, \( \bar{y} \) is the average value of the predicted value sequence and \( \bar{y}' \) is the average value of the real value sequence.

### Table 1. Measure indicators.

| Name  | Description          | Formula                                                                 |
|-------|----------------------|-------------------------------------------------------------------------|
| MSE   | mean squared error   | \( MSE = \frac{\sum_{t=1}^{N} (y(t) - \hat{y}(t))^2}{N - 1} \)          |
| SCC   | Squared correlation coefficient | \( SCC = \frac{\left( \sum_{t=1}^{N} y(t) \hat{y}(t) - N\bar{y}' \bar{y} \right)^2}{\left( \sum_{t=1}^{N} y(t) - N\bar{y} \right) \left( \sum_{t=1}^{N} \hat{y}(t) - N\bar{y}' \right)^2} \) |
| RMSE  | root mean squared error | \( RMSE = \sqrt{\frac{\sum_{t=1}^{N} (y(t) - \hat{y}(t))^2}{N - 1}} \)     |
| R     | correlation coefficient | \( R = \frac{\sum_{t=1}^{N} (y(t) - \bar{y})(\hat{y}(t) - \bar{y})}{\sqrt{\sum_{t=1}^{N} (y(t) - \bar{y})^2 \sum_{t=1}^{N} (\hat{y}(t) - \bar{y})^2}} \) |
| MAE   | mean absolute error  | \( MAE = \frac{1}{N} \sum_{t=1}^{N} |y(t) - \hat{y}(t)| \)          |

3.2. Experiments on Hang Seng Index
For SVM, there are some parameters which need to be chosen, including kernel function type \( t \), penalty parameter \( c \), kernel function parameter \( g \) and loss function \( p \). Experiments show that when the value of \( t \) is 0, the effect is better than other kernel functions, that is, linear kernel function is the best. \( p \) is determined by repeated experiments and the cross validation method is used to find the optimal parameters \( c \) and \( g \). For the prediction of the Hang Seng Index, the parameters of SVM of all algorithms are as shown in Table 2.

### Table 2. The parameters of SVM on the prediction of Hang Seng Index

| Algorithm | \( t \) | \( C \) | \( g \)    | \( p \)   |
|-----------|--------|--------|----------|----------|
| ENW-MPCA  | 0      | 500    | 0.0423   | 0.00001  |
| MPCA      | 0      | 450    | 0.07458  | 0.00001  |
| KPCA      | 0      | 1000   | 0.01088  | 0.00001  |
When using ENW-MPCA and MPCA algorithms for the feature extraction of Hang Seng Index, \( testQ \), which is the ratio of eigenvalues and sum of eigenvalues, is an important parameter. It determines the dimension of projection matrix of each mode, and has a strong influence on the prediction results. The dimension of tensor data is 20*34*7. To determine the best projection space, change the value of \( testQ \), and Figure 2 shows the curves of the SCC versus the variation of \( testQ \). From Figure 1, ENW-MPCA and MPCA algorithm are basically increasing with the increase of \( testQ \). Curves do not change in the former period, for the reason that the dimensions of features retained remain the same, which causes the changeless of SCC. When \( testQ \) increases to 0.9994, the feature tensor extracted is 14*7*6 and SCC of both algorithms reach the maximum. SCC for ENW-MPCA is 0.9133, and for MPCA is 0.9058. But when \( testQ \) continues to increase, SCC of the two algorithms fall instead, because the increase of redundancy leads to the decrease of prediction accuracy.

Figure 2 shows the closing price prediction of Hang Seng Index based on ENW-MPCA, MPCA, KPCA and PCA algorithms. As can be seen from Figure 3, comparing ENW-MPCA with MPCA, the change of the two curves are similar, but ENW-MPCA has better agreement, smaller fluctuation and smoother curve. MPCA agrees better than KPCA and PCA, and the results are better than the two algorithms, which shows that the feature extraction algorithm based on tensor is better than algorithms based on vector.
When achieving the optimal prediction accuracy, measure indicators on Hang Seng Index by the four algorithms are shown in Tab.4. Compared with MPCA, KPCA and PCA, the performance of ENW-MPCA is optimal, and the prediction accuracy is the highest. MSE of ENW-MPCA is 69050, and the SCC is 0.9136, better than 77783 and 0.9065 of MPCA, which indicates that its predictive effect is better. This shows that using the normalized eigenvalue as the weight of corresponding eigenvector can effectively distinguish the role of different eigenvalues and has a good inhibitory effect on the inherent redundancy information extraction of MPCA algorithm, which makes a better prediction accuracy. But the error limit is -852, error of the twenty-seventh trading days, so the error range of ENW-MPCA is from -900 to 600, the scope is slightly larger than that of MPCA, which is from -800 to 600.

The predictive effect of MPCA algorithm is better than KPCA and PCA, and is more stable. This shows that MPCA, based on tensor, can better retain the original information of data, and has better prediction performance. The measure indicators of KPCA are superior to PCA, which is because the introduction of the kernel function makes KPCA enable to consider the nonlinear correlation between sample data, so the error is small.

| Algorithm   | MSE    | SCC     | RMSE   | R     | MAE    | Error range |
|-------------|--------|---------|--------|-------|--------|-------------|
| ENW-MPCA    | 69050  | 0.9136  | 262.774| 0.9558| 208.1597| [-900,600]  |
| MPCA        | 77783  | 0.9065  | 278.896| 0.9521| 227.6137| [-800,600]  |
| KPCA        | 85708  | 0.8977  | 292.759| 0.9474| 234.7234| [-900,700]  |
| PCA         | 95825  | 0.8943  | 309.556| 0.9457| 252.6358| [-1000,700]|

The selection of feature dimension is also related to the prediction accuracy and efficiency. The relationship between the square correlation coefficient and the feature dimension of the prediction result is analyzed. Figure 3 shows the curve of the square correlation coefficient of Hang Seng Index with the change of feature dimension. For ENW-MPCA, MPCA, KPCA and PCA, when the feature dimension alters, SCC will be different. This indicates that the chosen of feature dimension can affect the discriminant feature of the four algorithms. From the experimental results, we can see that when the feature dimension is very small, the value of SCC is slow, which is because a small number of features cannot retain enough information of data. SCC of the four algorithms increases with the increase of feature dimension firstly, then increases slowly, and then decreases after the highest point is reached. But in the whole process, the movement trend is slightly fluctuation and not smooth enough. Compared with MPCA, SCC of ENW-MPCA fluctuates more slowly with the increase of feature dimension, and the maximum of it is higher than that of the MPCA algorithm.

![Figure 3. SCC versus feature dimension on Hang Seng Index](image)
4. Conclusion
This paper puts forward a prediction method of stock market index time series based on ENW-MPCA algorithm, using tensor method to predict stock price. Making no distinction, MPCA algorithm equally treats the role of different eigenvalues, without reflecting the different influences of the major and minor eigenvalues on the eigenvectors. Therefore, improve the MPCA algorithm, using the eigenvalues after normalizing as weight of the corresponding eigenvector, and ENW-MPCA algorithm is proposed. A three-dimensional tensor model is constructed by using the historical data of the stock market indexes. After adopting ENW-MPCA algorithm to reduce the dimension and extract the features, the support vector machine is used for the regression prediction, and finally the prediction results are obtained.

Experimental results show that compared with MPCA, ENW-MPCA has better predictive effect and higher accuracy. Compared with KPCA and PCA, which are based on vector, the performance of feature extraction of ENW-MPCA is better. This shows the relationship between the structure of internal information and data retention data algorithm based on tensor can better retain the structure information of internal data the relationship between data, showing a better prediction performance. This paper verifies the feasibility of applying tensor model to predict the stock market index time series. With the development of stock market in China, it is believed that the proposed algorithm can be applied in a wide range of fields and areas.

Acknowledgments
This work was financially supported by the The National Natural Science Foundation of China (51879211).

References
[1] Rubio, A., Bermúdez, J.D., Vercher, E. (2017) Improving stock index forecasts by using a new weighted fuzzy-trend time series method. Expert Systems with Applications, 76:12–20.
[2] Oriani, F.B., Coelho, G.P. (2016) Evaluating the impact of technical indicators on stock forecasting. In: 2016 IEEE Symposium Series on Computational Intelligence. Athens. pp.1-8.
[3] Rout, A.K., Bisoi, R., Dash, P.K. (2015) A low complexity evolutionary computationally efficient recurrent Functional link Neural Network for time series forecasting. In: 2015 IEEE Power, Communication and Information Technology Conference. Bhubaneswar. pp.576-582.
[4] Yu, J., Kim, S. (2015) Automatic structure identification of TSK fuzzy model for stock index forecasting. In: 2015 IEEE International Conference on Fuzzy Systems. Istanbul. pp.1-8.
[5] Gholamiangonabadi, D., Taheri, S.D.M., Mohammadi, A., Menhaj, M. B. (2014) Investigating the performance of technical indicators in electrical industry in Tehran's Stock Exchange using hybrid methods of SRA, PCA and Neural Networks. In: 2014 5th Conference on Thermal Power Plants. Tehran. pp.75-82.
[6] Grigoryan, H.(2016) A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA). Database Systems Journal, 7(1):12-21.
[7] Guo, Z., Wang, H., Liu, Q., Yang, J. (2014) A Feature Fusion Based Forecasting Model for Financial Time Series. Plos One, 2014, 9(6):172-200.
[8] Guo, Z., Zeng, Y., Yang, J., Ye,W. (2017) Time Series Forecasting of Stock Market Index based on MPCA-RBF Model. Application Research of Computers, 2017(11):1-7.