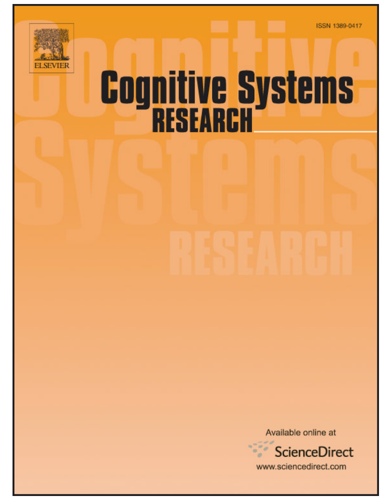
# Accepted Manuscript




Index Tracking Based on Deep Neural Network

Hongbing Ouyang, Xiaowei Zhang, Hongju Yan

PII: S1389-0417(18)30439-X
DOI: https://doi.org/10.1016/j.cogsys.2018.10.022
Reference: COGSYS 763

To appear in: *Cognitive Systems Research*

Received Date: 1 August 2018
Revised Date: 3 October 2018
Accepted Date: 11 October 2018

Please cite this article as: Ouyang, H., Zhang, X., Yan, H., Index Tracking Based on Deep Neural Network, *Cognitive Systems Research* (2018), doi: https://doi.org/10.1016/j.cogsys.2018.10.022

# Index Tracking Based on Deep Neural Network

Hongbing Ouyang, Xiaowei Zhang, Hongju Yan*
*School of Economics, Huazhong University of Science and Technology, China*
*Corresponding Author's Email:* qfnuyhj@163.com

**Abstract:** Deep learning has a strong ability to extract feature representations from data, since it has a great advantage in processing nonlinear and non-stationary data and reflecting nonlinear interactive relationship. This paper proposes to apply deep learning algorithms including deep neural network and deep autoencoder to track index performance and introduces a dynamic weight calculation method to measure the direct effects of the stocks on index. The empirical study takes historical data of Hang Seng Index (HSI) and its constituents to analyze the effectiveness and practicability of the index tracking method. The results show that the index tracking method based on deep neural network has a smaller tracking error, and thus can effectively track the index.

**Keywords:** deep neural network; deep autoencoder; index tracking

## 1.Introduction

Index tracking requires to construct and manage a portfolio of stocks whose performance is as close as possible to that of a given stock index [1]. Compared with other mutual funds, index fund is a relatively stable and effective investment tool, which is very important in modern portfolio management practice. The full replication method of index portfolio management is to incorporate all constituents of the index into the fund, which would lead to high transaction costs and greater tracking errors. Therefore, it is of great significance for developing new approaches and providing decision-making reference for investors to explore how to replicate the index with a relatively small number of stocks, reduce tracking errors and management costs and even achieve better performance than the index.

Based on Markowitz's mean-variance model, many scholars study on the optimized replication of index by minimizing the volatility of tracking error [2], which is also known as the early study on index portfolio management. Following studies [3,4] consider different definition of tracking error and tried different replica methods based on the framework proposed by [2]. These traditional index tracking management methods studied by [2,5,6,7] provide an effective tool for index fund management and a feasible analytical framework for index tracking research.

However, with the development of machine learning, genetic algorithm (GA) and other artificial intelligence (AI) algorithms, a lot of meaningful explorations on index tracking based on these AI techniques have been made. Oh et al. [8] propose a portfolio optimization framework for index fund management based on GA. In this approach, the stocks are selected through fundamental analysis, and the weights of the selected stocks are selected are optimized by GA. Empirical results show that the strategy proposed by [8] is more practical than conventional portfolio mechanism. Beasley et al. [9] present an evolutionary heuristic method for the index tracking problem, the results show that it's a much more desirable proposition than tracking index by full replication. Fernandez and Gomez [10] propose a heuristic method based on Hopfield neural network is used it to solve the general mean-variance portfolio selection model, but the experimental results don't show any outperformance with other methods. Ellis and Wilson [11] apply neural network to the property sector stocks to construct value portfolios, concluding that neural network can construct property portfolios that outperform the market on a regular basis. Freitas et al. [12] present a prediction-based portfolio optimization model using neural network that can capture short-term investment opportunities, simulation results show that this method outperforms the mean-variance model and beats the market index. So, the existing research methods for index portfolio management (i.e., index tracking) based on BP neural network [10,11,12,13], support vector machine [14,15] and so on achieve better performance compared to other traditional methods.

In recent years, the success of deep learning in many disciplines in AI fields such as AlphaGo, autonomous vehicles, speech recognition and pattern recognition has attracted the attention of more and more scholars. Deep learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level into a representation at a higher, slightly more abstract level, with the composition of enough such transforms, very complex functions can be learned [16]. Moreover, data representation is empirically found to be a core determinant of the performance of most machine learning algorithms, while deep learning can yield more non-linear, more abstract representations [17].

Therefore, it is of great theoretical and practical significance to apply deep learning algorithms in index tracking, so as to capture the nonlinear and non-stationary features in financial stock time series data, reflect the nonlinear interactive relationship between different stocks and reduce the tracking error. A deep portfolio construction framework based on deep learning [18]. However, it didn't discuss the determination of stock weights then the direct effect of stocks on the index cannot be obtained.

Following the framework proposed by [18], this paper first explores how to select stocks based on deep autoencoder combined with capital asset pricing model (CAPM), and then studies how to fit the selected stocks and the index based on deep neural network. Most importantly, different from [18], this paper explores the similarity between deep autoencoder and Capital Asset Pricing Model and proposes a dynamic weight calculation method to determine the respective stock weight for the index. It is of great significance to propose the dynamic weight calculation method both in theory and application for it not only provides a simple way to test the effectiveness of the index tracking but also a feasibility to apply the proposed index tracking method in the practical index tracking.

The following structure of this paper is as follows: Part 2 is the construction of the index tracking method based on deep learning, Part 3 is the empirical analysis, where the index tracking method is applied to the analysis of Hang Seng Index (HSI) and its' constituent stocks and Part 4 is the conclusion.

## 2. Methodology

This section first discusses how to select a relatively small number of stocks from the constituent stocks of the target index based on deep autoencoder, and then explores how to use the deep neural network to fit the selected stocks with the index. Also, this section proposes the dynamic weight calculation method to track the index efficiently.

### 2.1 Deep autoencoder

The deep autoencoder can be regarded as the nonlinear generalization of Principal Component Analysis (PCA). The multi-level encoder network converts high-dimensional data into low-dimensional encoded data, and similar decoder network reconstructs low-dimensional encoded data to obtain output information. Compared with the autoencoder, the deep autoencoder has better compression efficiency [19]. The structure of the deep autoencoder is shown in Figure 1.

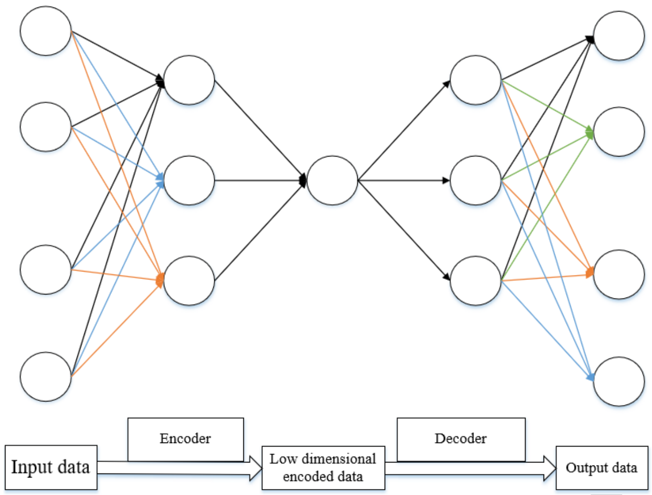



**Figure 1.** The Structure of Deep Autoencoder

Preset the input data vector as $X=[x_1,x_2,...,x_n]$, where $x_i(i=1,2,...,n)$ is the time series data of the *i*th stock. The output data vector is $X'=[x_1^{'},x_2^{'},...,x_n^{'}]$, and the low-dimensional encoded data is *Z*. The encoder process is:

$$Z=f_2(W_2f_1(W_1X+b_1)+b_2) \tag{1}$$

And the decoder process is:

$$X'=f_4(W_4f_3(W_3Z+b_3)+b_4) \tag{2}$$

where, $W_1,W_2,W_3,W_4$ are neural network weights, $b_1,b_2,b_3,b_4$ are biases, and $f_1$, $f_2$, $f_3$, $f_4$ are the activation functions, which can realize nonlinear transformation for weighted data. Autoencoder process with nonlinear encoder functions $f_1$, $f_2$ and nonlinear decoder functions $f_3$, $f_4$ can thus learn a more powerful nonlinear generalization of PCA.

The low-dimensional encoded data *Z* generated by the encoder process equation (1) can be considered as the market portfolio composed of *n* stocks and contains the nonlinear interactive effect between different stocks. The output data $X'$ (i.e., the data after decoding) generated by equation (2) reflects the nonlinear impact of market portfolio on different stocks.

It is important to note that this deep autoencoder process shares certain similarities with Capital Asset Pricing Model (CAPM) proposed by [20,21]. The similarity between them is that both of low-dimensional encoded data *Z* and market portfolio $R_m$ are composed of different stocks in the financial market. The difference between them is that the deep autoencoder process as shown in Figure 1 reflects the nonlinear interaction among stocks and the effect of low-dimensional encoded data on stock is nonlinear, while the CAPM assumes that the market portfolio is constructed by the linearly weighted method and the impact of market portfolio on stock is linear.

So, we assume that the decoded data $X'$ contains the market portfolio information. The similarity between the reconstructed stock data $X'$ and the original data *X* can be used to measure the common information between stock and market portfolio information. Following [18], this paper proposes equation (3) to measure the common information of stocks and market portfolio:

$$d_i = \| x_i - x_i' \|^2 \tag{3}$$

The smaller the value of $d_i$ is, the more common information contains between stocks and market portfolio. There's no benefit in having multiple stocks contributing the same information [22]. Therefore, this paper selects *h* stocks that share much common information with market portfolio and *l* ($h < l$) stocks that share less common information with market portfolio through the deep autoencoder to track the market portfolio index.

Therefore, based on the deep autoencoder process, the stocks of index tracking portfolio are chosen preliminarily, which is called the stock selecting process. To this end, this paper needs to further estimate the parameters in equation (1) and equation (2), optimize the function based on the autoencoder structure and activation function to minimize the error between the original data and reconstructed data, so as to determine the optimal parameter of $W_1, W_2, W_3, W_4$ and $b_1, b_2, b_3, b_4$, that is:

$$\min |X - X'| + \lambda J \tag{4}$$

Where, the regularization term $J = \frac{1}{2}\|W\|_2^2 = \frac{1}{2}W^T W$ limits the model's learning ability and improve its generalization ability, $\lambda \in [0, +\infty)$ measures the parameter penalty term *J* and the hyper-parameter of the relative contribution of the standard target function $|X - X'|$.

### 2.2 Deep neural network and measurement of index weight

After selecting stocks, this paper will further consider how to effectively track the index based on these stocks. The construction of investment index tracking based on deep neural network has the following advantages: Firstly, the multiple processing layers of deep neural networks can effectively extract representations of the time series change of each stock. Secondly, the index tracking based on deep neural network can reflect the nonlinear interaction of different stocks. Therefore, this paper uses the deep neural network to track the index, and then determine the weight of each stock.

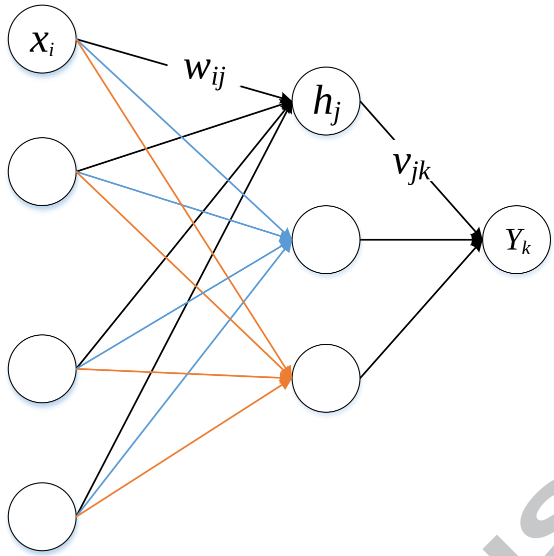



**Figure 2.** The structure of deep neural network

Without loss of generality, this paper discusses the fundamental principle of the deep neural network and the measurement of corresponding index weight based on the deep neural network structure as shown in Figure 2. Preset the input data vector as $X=[x_1, x_2, ..., x_n]$, $x_i(i=1,2,...,n)$ is the time series data of the $i$th stock. The output of the hidden layer is $H=[h_1, h_2, ..., h_n]$, where $h_j(j=1,2,...,m)$ is the output of the $j$th hidden unit. If the output of the output layer is $Y_k$, $\sigma_1$ and $\sigma_2$ represent the corresponding activation functions in the hidden and output layers, respectively. We get the following relational expression:

$$out_j=\sum_{i=1}^{n} x_i w_{ij} \quad , \quad h_j=\sigma_1(out_j) \tag{5}$$

$$out_k=\sum_{j=1}^{m} y_j v_{jk}+b_k, \quad Y_k=\sigma_2(out_k) \tag{6}$$

It should be noted that nonlinear activation function can realize the nonlinear transformation of the weighted data. So, in equation (5) and (6), nonlinear functions $\sigma_1$ and $\sigma_2$ can reflect the nonlinear interaction relationship among different stocks and nonlinear influence of stocks on index.

Equation (7) and Equation (8) are the vector expression of equation (5) and (6):

$$H=\sigma_1(XW+b_1) \tag{7}$$

$$Y_k=\sigma_2(HV+b_2) \tag{8}$$

where $W = \begin{bmatrix} w_{11}, w_{12}, \ldots, w_{1m} \\ w_{21}, w_{22}, \ldots, w_{2m} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ w_{n1}, w_{n2}, \ldots, w_{nm} \end{bmatrix}, V = \begin{bmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{n1} \end{bmatrix}$.

The parameters of neural network can be obtained through the following optimization function similar with equation (4):

$$\min \left| Y_k \text{-} \hat{Y}_k(X) \right| + \lambda J(\theta) \tag{9}$$

where $J(\theta) = \frac{1}{2}(\|W\|_2^2 + \|V\|_2^2)$ is the regularization term.

The weight matrixes $W$ and $V$ trained through the deep neural network reflect the relationship between different units in different layers of the neural network. But they cannot reflect the influence of stock on index directly, and thus cannot determine the weights of stocks in the constructed portfolios. So, this paper further proposes a method to calculate the weights of input unit towards output unit to capture the direct relationship between input factor and output factor. That is:

$$\begin{aligned} \frac{\partial Y_k}{\partial x_i} &= \sigma_2(HV)'\sigma_1(XW)'V\frac{\partial(XW)}{\partial x_i} \\ &= \sigma_2(HV)'\sigma_1(XW)'VW_i \end{aligned} \tag{10}$$

where $W_i = [w_{i1}, w_{i2}, \ldots, w_{im}]$ $(i = 1, 2, \ldots, n)$ . It is important to note that $\sigma_2(HV)' = 1$ and $\sigma_1(XW)' = 1$ when the activation functions $\sigma_1$ and $\sigma_2$ are linear activation functions. In this case, $\partial Y_k / \partial x_i$ is a time-independent constant. If $\sigma_1$ and $\sigma_2$ are nonlinear activation functions, $\partial Y_k / \partial x_{it}$ of different time may vary, which is often neglected in existing literature.

## 3.Empirical analysis

For empirical study, this paper uses HSI and its constituent stocks data to show the effectiveness of the index tracking method constructed in this paper.

### 3.1 Data description and processing

By the end of January 10, 2018, HSI contains 51 constituent stocks. This paper uses the weekly closing price of HSI and its 46 constituent stocks (excluding 1928.HK, 1113.HK、1299.HK, 0288.HK and 1997.HK) from January 2, 2009 to December 29, 2017 to construct the investment index portfolio due to limited data availability. The sample is divided into three subsets: training set, validation set and test set. The details are shown in Table 1. The training set is used to train model parameters. The validation set is used to adjust the deep autoencoder and deep neural network structure as well as parameters. The test set is used to evaluate the effectiveness of the index portfolio model.

Table 1 Division of Sample Data

| Data | Sample size | Time |
|---|---|---|
| Training set | 352 | 2009/01/02-2015/09/25 |
| Validation set | 59 | 2015/10/02-2016/11/11 |
| Testing set | 59 | 2016/11/18-2017/12/29 |

The standardized processing of characteristics of sample data plays an important role for the optimal effect of deep learning algorithm. This paper standardizes all the time series data. For time series $x_1, x_2, \dots, x_t$, the standardization is as follows:

$$y_i = \frac{x_i - \overline{x}}{s}, \quad i = 1, 2, \dots, t \tag{11}$$

Where $\overline{x} = \frac{\sum x_i}{t}$, $s = \sqrt{\frac{1}{t-1}(\sum x_i - \overline{x})^2}$.

### 3.2 Stock selection

Deep autoencoder is used to explore the common information between 46 constituent stocks of HSI and market portfolio, and then determining the stocks required for the construction of investment portfolio. Figure 3 and Figure 4 give the stock (2382.HK) that shares the most common information with market portfolio and the stock (0762.HK) that shares the least common information with market portfolio separately. It's clear that the autoencoder reconstruction data of stock 2382.HK fit with the original data much better than the stock 0762.HK.

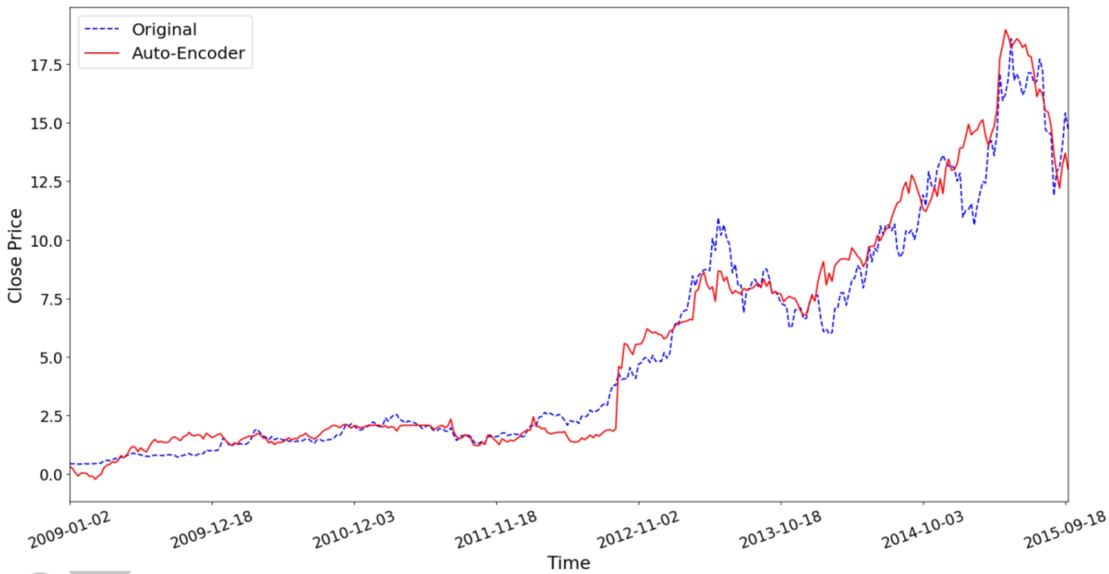



**Figure 3.** The Stock shares the most common information with market portfolio (2382.HK)

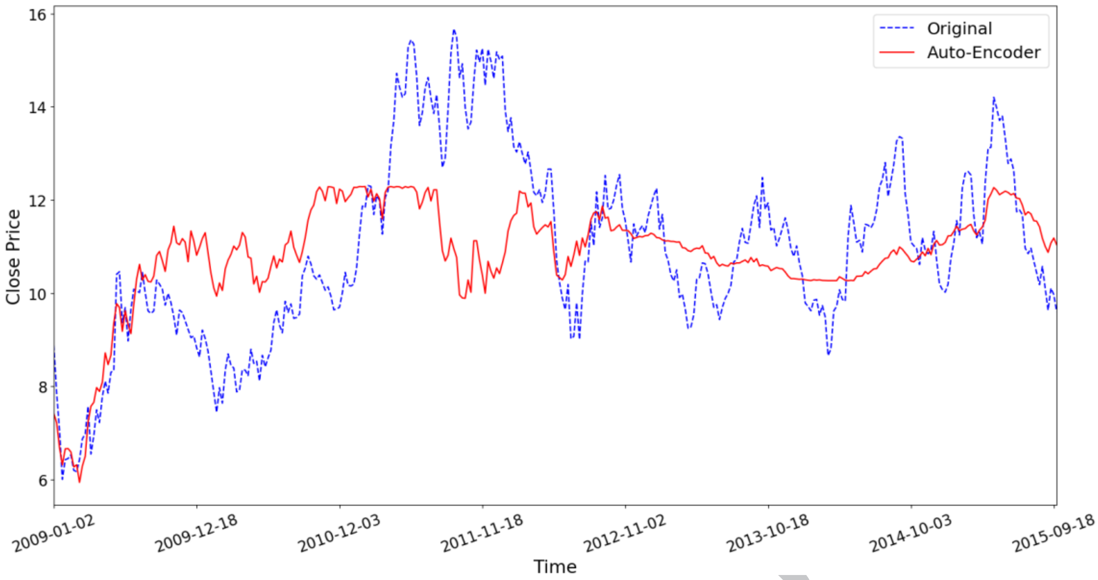



**Figure 4.** The Stock shares the least common information with market portfolio (0762.HK)

Given the fact that the selection of multiple stocks doesn't make much sense for reflecting the same information, this paper selects 3 stocks (2382.HK, 0700.HK, 0175.HK) that share more common informaiton with the market and 5 stocks (0267.HK, 3328.HK, 0017.HK, 2628.HK, 0762.HK) that share less common information with the market to construct the investment index portfolio which tracks HSI.

### 3.3 Index portfolio Construction

The stocks selected by the deep autoencoder are used to construct the tracking portfolio based on deep neural network. The tracking effect of portfolio construction in the training set and validation set are shown in Figure 5 and Figure 6 respectively.

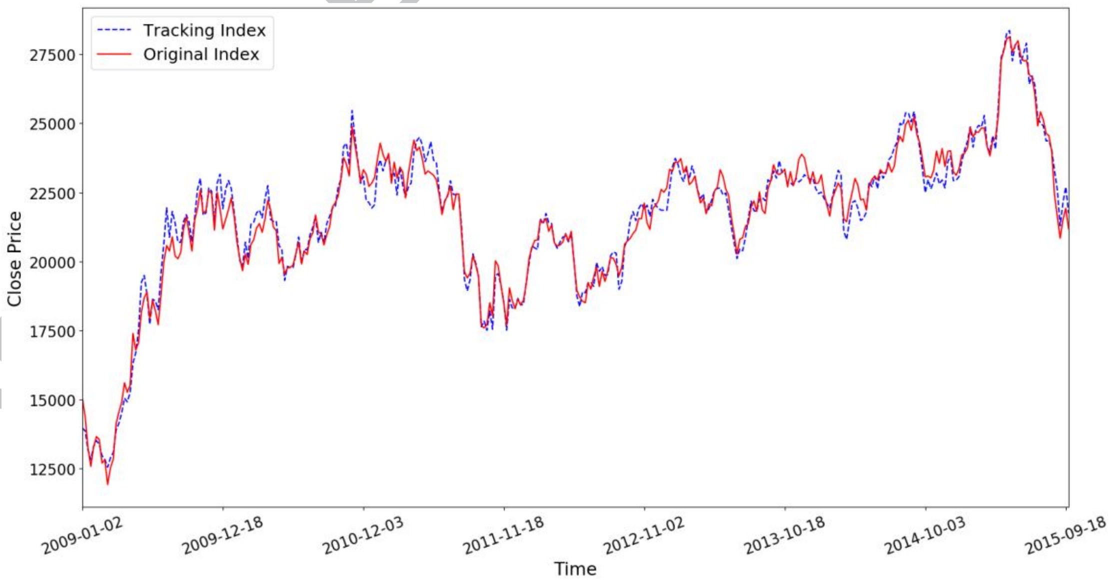



**Figure 5.** Tracking effect (Training set)

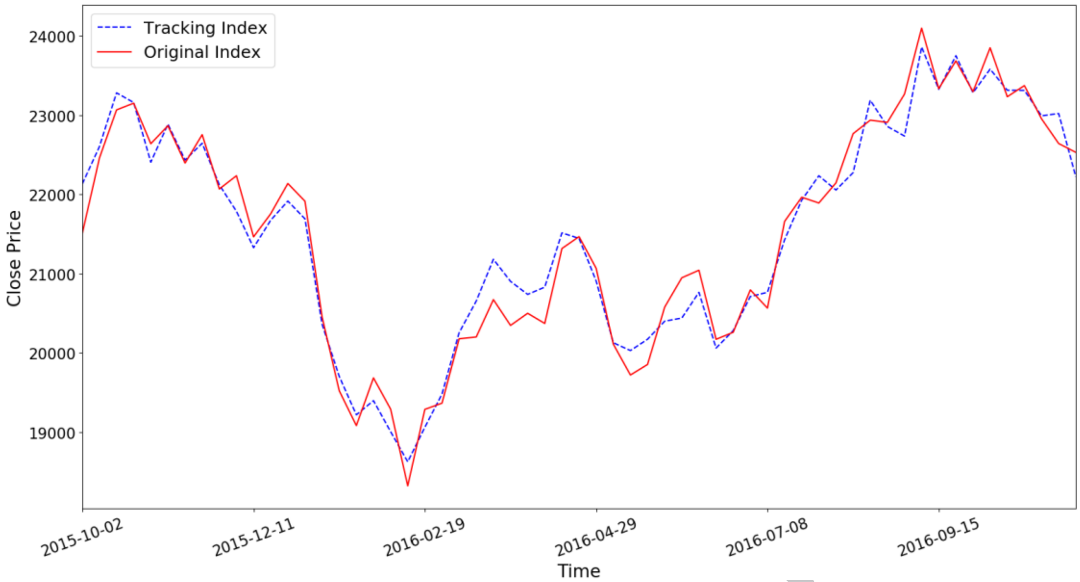



**Figure 6**. Tracking effect (Validation set)

It can be found that the index tracking portfolio constructed in this paper can effectively track the index both in training and validation sets. Thus, this method could be used to effectively track the trends of HSI. To further explore its tracking performance in test set, the measurement of the weights of each stock based on equation (10) mentioned above is needed, that is,

$$\begin{aligned}\frac{\partial Y}{\partial x_i} &= \sigma_2(HV)'\sigma_1(XW)'V\frac{\partial(XW)}{\partial x_i} \\ &= \sigma_2(HV)'\sigma_1(XW)'VW_i\end{aligned} \tag{12}$$

The activation functions in the hidden layer and output layer adopted in this paper are ReLU activation function (that is, $\sigma_1(x) = \max(0, x)$ ) and Linear activation function (that is, $\sigma_2(x) = x$ ) respectively. So:

$$\sigma_1(x)' = \begin{cases} 0 & x \le 0 \\ 1 & x > 0 \end{cases}, \quad \sigma_2(x)' = 1 \tag{13}$$

Therefore, the weight of corresponding stocks at different time may vary, that is, the impact of the same stock on index portfolio may vary at different time. Equation (12) and equation (13) can be used to figure out the weight of each stock in tracking the index. The weights of the stocks are rounded off because the minimum number of a trade in Hong Kong Stock Exchange is one round lot. It should be noted that 1 round lot of 2383.HK and 0700.HK contain 100 shares, while others contain 1000 shares.

According to the weights calculated, an index portfolio tracking method including long and short equity mechanism is generated. For example, in 18 November, 2016, the index portfolio tracking method is constructed as follows: buy 20 round lots 0700.HK, 15 round lots 0175.HK, 14 round lots 3328.HK, 1 round lot 0017.HK and short 70 round lots 2382.HK, 3 round lots 0267.HK, 11 round lots 0762.HK. Since the weight of stocks is different in each week, the index portfolio is adjusted every week. Consequently, the tracking effect of out-of-sample index portfolio could be obtained, which is shown in Figure 7. It can be seen from Figure 7 that the index portfolio based on deep autoencoder and deep neural network can effectively track the trends of HSI. Figure 7 shows that using deep autoencoder and deep neural network could track the index through less stock and reducing transaction costs, which is of great importance in theory and practice.

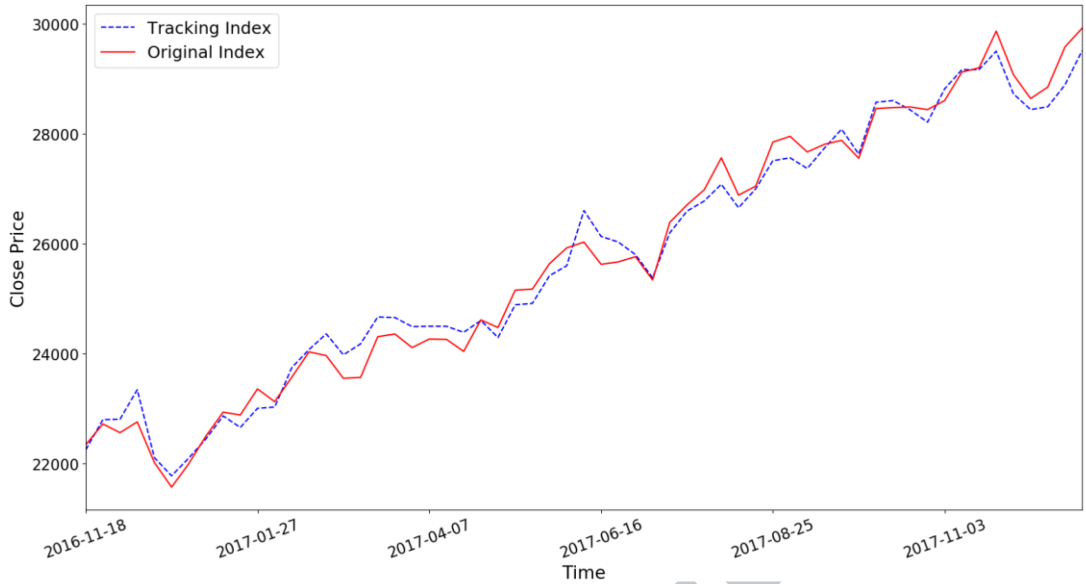



**Figure 7.** Tracking effect (Testing set)

To further confirm the effectiveness of the method, we calculates the tracking error and cumulative return to explore the tracking effect. The tracking error refers to the proximity of index portfolio to its original index, that is,

$$TE = \frac{1}{T}\sqrt{\sum_{t=1}^{T}(R_{It} - R_{pt})^2} \tag{14}$$

Where $R_{Pt}$ and $R_{It}$ are the return of index portfolio and original index in time t respectively. The smaller TE is (TE>0), the more effective the method is. The weekly tracking error TE=0.12%. Compared to the existing tracking methods [23,24], that of proposed in this paper have smaller tracking error.

Table 8 shows the cumulative return of the tracking index and the original index. The cumulative returns of index portfolio and original index from November 18, 2016 to December 29, 2017 are 29.10% and 30.01% respectively. This reaffirms that index portfolio based on deep learning shows good performance on tracking the original index.

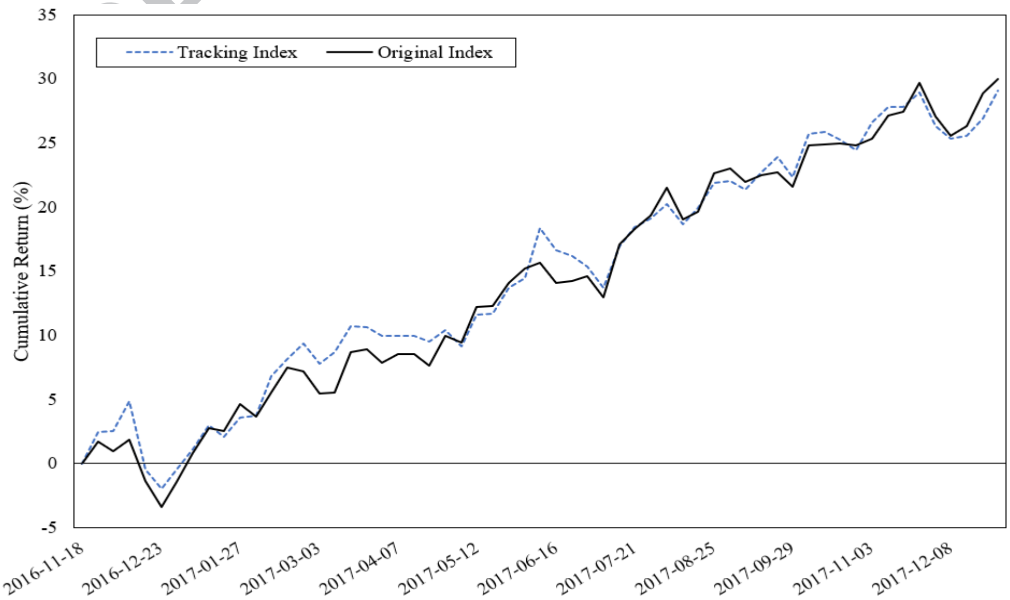


**Figure 8.** Cumulative Return

## 4. Conclusion and discussion

This paper adopts deep autoencoder and deep neural network to track the index by using the weekly data of HSI and its constituent stocks. The deep autoencoder is used for selecting stocks while the deep neural network is built for constructing the index portfolio. The empirical results prove the effectiveness of the method in tracking index. Different from the existing studies, this paper firstly discusses the similarity between deep autoencoder and capital asset pricing model (CAPM) in reflecting the market information as well as the advantages of deep autoencoder in measuring the common information between individual stocks and the market portfolio. Meanwhile, this paper introduces a dynamic weight calculation method to measure the effects of stocks on the index directly, which is not done before. It has important significance to extend the applicability of index tracking method based on deep neural network both in theory and practice.

It is noteworthy that the sequence correlation of financial series cannot be reflected completely by deep neural network. There's still some limitation to fit the selected stocks and index based on deep neural network. So, further studies will be needed to explore how to capture the financial time series dependence and thus improves the tracking accuracy.

**Acknowledgement**

The authors acknowledge the funding from Fundamental Research Funds for the Central Universities (Grant No.2018WKZDJC004) and the support from Nanning Development and Reform Committee project "Construction of Nanning Channel Evaluation Index System and its Application".

**References**

1. Corielli, F., & Marcellino, M. (2006). Factor based index tracking. Journal of Banking & Finance, 30(8), 2215-2233.
2. Roll, R. (1992). A mean/variance analysis of tracking error. Journal of Portfolio Management, 18(4), 13-22.
3. Rudolf, M., Wolter, H. J., & Zimmermann, H. (1999). A linear model for tracking error minimization. Journal of Banking & Finance, 23(1), 85-103.
4. EI-Hassan, N., & Kofman, P. (2014). Tracking error and active portfolio management. Australian Journal of Management, 19(2), 267-76.
5. Jansen, R., & Van Dijk, R. (2002). Optimal benchmark tracking with small portfolios. The Journal of Portfolio Management, 28(2), 33-39.
6. Frino, A., Gallagher, D., Neubert, A., & Oetomo, T. (2004). Index design and implications for index tracking: Evidence from S&P 500 Index Funds. Journal of Portfolio Management, 30(2), 89-95.
7. van Montfort, K., Visser, E., & van Draat, L. F. (2008). Index tracking by means of optimized sampling. Journal of Portfolio Management, 34(2), 143.
8. Oh, K. J., Kim, T. Y., & Min, S. (2005). Using genetic algorithm to support portfolio optimization for index fund management. Expert Systems with Applications, 28(2), 371-379.
9. Beasley, J. E., Meade, N., & Chang, T. J. (2003). An evolutionary heuristic for the index tracking

problem. European Journal of Operational Research, 148(3), 621-643.
10. Fernandez, A., & Gomez, S. (2007). Portfolio selection using neural networks. Computers & Operations Research, 34(4), 1177-1191.
11. Ellis, C., & Wilson, P. (2005). Can a neural network property portfolio selection process outperform the property market? Journal of Real Estate Portfolio Management, 11(2), 105-121.
12. Freitas, F. D., De Souza, A. F., & de Almeida, A. R. (2009). Prediction-based portfolio optimization model using neural networks. Neurocomputing, 72(10-12), 2155-2170.
13. Yu, L., Wang, S., & Lai, K. K. (2008). Neural network-based mean–variance–skewness model for portfolio selection. Computers & Operations Research, 35(1), 34-46.
14. Hu, C. P., Xue, H. G., & Xu, F, M. (2014). A stock index replicating model based on time weighted svm and it's empirical analysis. Systems Engineering-Theory & Practice, 34(9), 2193-2201.
15. Yang, G. L., Zhao, S. T., Xu, C. X. (2009). A research on svm based financial market index tracking technology. Studies of International Finance, 10, 68-72.
16. Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature,521(7553), 436.
17. Bengio, Y., Courville, A., & Vincent, P. (2012). Unsupervised feature learning and deep learning: a review and new perspectives.
18. Heaton, J. B., Polson, N. G., & Witte, J. H. (2016). Deep learning for finance: deep portfolios. Applied Stochastic Models in Business & Industry, 33(1).
19. Hinton G E, Salakhutdinov R R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786):504.
20. Sharpe, W. F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. Journal of Finance, 19(3), 425-442.
21. Lintner J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. The review of economics and statistics, 13-37.
22. Heaton, J. B., Polson, N. G., & Witte, J. H. (2016b). Deep portfolio theory. Papers.
23. Cai, Y. P., Wan, L, Fan, X. D. (2006). An empirical comparison of alternative models of index fund. Journal of Quantitative & Technical Economics, 23(10), 130-140.
24. Jian, G., Yang D. (2006). A comparative analysis of replication method of index investment. Journal of Financial Research,51(8), 31-40.