Scientific Research Publishing

Journal of Mathematical Finance, 2019, 9, 266-271
http://www.scirp.org/journal/jmf
ISSN Online: 2162-2442
ISSN Print: 2162-2434

# Possibility for Short-Term Forecasting of Japanese Stocks Return by Randomly Distributed Embedding Theory

**Seisuke Sugitomo[1], Keiichi Maeta[2]**

[1]Fund Manager at Epic Partners Investment Co., Ltd., Tokyo, Japan
[2]Graduate School of Mathematical Sciences, University of Tokyo, Tokyo, Japan
Email: sugitomo@epicgroup.jp, maeta@ms.u-tokyo.ac.jp

**How to cite this paper:** Sugitomo, S. and Maeta, K. (2019) Possibility for Short-Term Forecasting of Japanese Stocks Return by Randomly Distributed Embedding Theory. *Journal of Mathematical Finance*, **9**, 266-271.
https://doi.org/10.4236/jmf.2019.93015

**Received:** May 8, 2019
**Accepted:** July 5, 2019
**Published:** July 8, 2019



## Abstract

In this work, we use the model-free framework, named randomly distributed embedding, which is the method that randomly selects variables from the values of many observed variables at a certain time and estimates the state of the attractor at that time, to predict the future return of Japanese stocks and show that the prediction accuracy is improved compared to the conventional methods such as simple linear regression or least absolute shrinkage and selection operator (LASSO) regression. In addition, important points to be considered when applying the randomly distributed embedding method to financial markets, and specific future practical applications will be presented.

## Keywords

Randomly Distributed Embedding (RDE), Least Absolute Shrinkage and Selection Operator (LASSO), Artificial Intelligence Finance, Japanese Stock Return Prediction

## 1. Introduction

For the portfolio management in the stock market, predicting accurately the return of stocks to be traded is an important issue. However, the prediction is not easy because financial data have a very low signal to noise ratio, the relationship between the data is intertwined complicatedly and it is difficult to obtain a sufficient number of samples in the time series.

On the other hand, among financial assets, the stock market has the characteristic that the number of stocks is very large and simultaneous measurement is possible although the amount of data in the time series direction is not large.

Therefore, it is considered that the randomly distributed embedding method (RDE) [1] has high affinity with the return prediction in the stock market. RDE is a mathematical framework to predict future changes of important target variables with high accuracy from the short-time series data consisting of simultaneous measurements of multiple variables, proposed in October 2018.

In this work, we will evaluate the effectiveness of the randomly distributed embedding method by comparing with the results of the methods using the simple linear regression and the least absolute shrinkage and selection operator (LASSO) regression.

The first author of this work, Seisuke Sugitomo, belongs to Epic Partners Investments Co., Ltd. The second author of this work, Keiichi Maeta, belongs to the Graduate School of Mathematical Sciences, the University of Tokyo.

## 2. Basic Concepts

### 2.1. Reconstruction of Attractors

We review the reconstruct theory according to [1].

Analysis of irregular time series signals observed in nature has been studied as "chaos time series analysis". In order to analyze irregular time series data from the viewpoint of deterministic dynamical systems, it is necessary to reconstruct the attractors [2], [3].

The most common method of attractor reconstruction is the reconstruction using the delay attractor.

The delay attractor is a reconstructed attractor of a dynamical system using the delay coordinate system $\left(x_k(t), x_k(t+\tau), x_k(t+2\tau), \cdots\right)$ with respect to a certain variable $x_k(t)$ ($t$ is time and $\tau$ is an interval.).

If the dimension of the delay coordinate system is larger than a certain level, there is an embedding $\Phi$ into the reconstructed attractor $M$ from the original attractor of the dynamical system according to Takens' embedding theorem [3] and the generalized embedding theorem [2].

On the other hand, the non-delay attractor is a reconstructed attractor of dynamical system using randomly select $m$ valuables from $\left\{x_i(t)\right\}$ ($m$ is the same number as the dimension of the delay coordinate system) and the coordinate system composed of them $\left(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_m}(t)\right)$. There is also an embedding $\Gamma$ from the original attractor to the reconstructed attractor $N$([2], [4], [5]).

### 2.2. Randomly Distributed Embedding Method

Randomly distributed embedding method is the method proposed by Aihara *et al.* [1] in October 2018 for predicting high-order, short-term time-series data with high accuracy.

First, we reconstruct the delay attractor and the non-delay attractor with respect to the observation data $x_i(t), (i=1,2,\cdots,n)$.

According to the embedding theory, there is a diffeomorphism $\Psi : M \to N$ compatible with embeddings $\Phi$ and $\Gamma$, and learning them from samples

enables a highly accurate prediction of the delay attractor using the non-delay attractor.

### 2.3. Application to Japanese Stocks

Next, we consider how to apply the above randomly distributed embedding method to the return prediction of Japanese stocks. The point to be noted in applying this method is that each variable in the observation data is a result from the same dynamical system.

Risk factors often used in the return prediction are unlikely to be attributed to the same dynamical system. On the other hand, each return of individual stocks included in the same industry is likely to be due to the same dynamical system.

So, in this work, we aim at predicting the return of a specific stock using the returns of individual stocks included in the same industry.

### 2.4. Gaussian Process Regression

Gaussian process regression is a nonparametric regression model [4]. Let us assume that the relation $t = w^{\mathrm{T}}\phi(x) + \varepsilon$ holds for two variables $x \in \mathbb{R}^n$, $t \in \mathbb{R}^n$ using basis functions

$$\phi : \mathbb{R}^n \to \mathbb{R}^n,$$

where $w \sim N(0, \alpha^{-1} I_n)$ and $\varepsilon \sim N(0, \beta^{-1} I_n)$ hold for a weight $w$ and an error $\varepsilon$.

At this time, we estimate the distribution of the output $t_{n+1}$, namely, $p(t_{n+1} \mid x_{n+1}, x_1, x_2, \cdots, x_n, t_1, t_2, \cdots, t_n) = N(t_{n+1} \mid m, \sigma^2)$ obtained from the test data $(x_1, t_1), (x_2, t_2), \cdots, (x_n, t_n) \in \mathbb{R}^n \times \mathbb{R}$ and the new input $x_{n+1}$.

We set the kernel function $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ $k(x, x') = \alpha^{-1}\phi(x)^{\mathrm{T}}\phi(x')$ and put $K = \left(k(x_i, x_j)\right)_{i,j}$, $t = (t_1, \cdots, t_n)^{\mathrm{T}}$, $k = \left(k(x_1, x_{n+1}), \cdots, k(x_n, x_{n+1})\right)^{\mathrm{T}}$.

Then, the optimal estimate is given by

$$m = k^{\mathrm{T}} K^{-1} t, \quad \sigma^2 = k(x_{n+1}, x_{n+1}) + \beta^{-1} - k^{\mathrm{T}} K^{-1} k. \tag{1}$$

## 3. Verification

### 3.1. Verification Procedure

The universe is TOPIX500 constitutive brand which is the top 500 stocks with high market capitalization and liquidity of the TOPIX adopted stocks. We apply the randomly distributed embedding method in the several industries using TSE 33 industry. In this work, seven types of industries, construction, chemistry, food, machinery, electronics, pharmaceuticals, and transportations are targeted for forecasting, because the number of stocks is within the industry to some extent and changes in the results of domestic and external demand are also to be examined. With regard to the randomly distributed embedding method, the prediction is performed according to the following procedure according to [1].

The given data is the data at time $t_1, \cdots, t_n$ of the function $x : \mathbb{R} \to \mathbb{R}^n$ $t \to (x_1, \cdots, x_n)$ at $n$ observation points. It is the next day daytime return of the

$k$-th specific variable $x_k$.

First, we choose s tuples containing $L$ numbers from $\{1,2,3,\cdots,n\}$. Then, from the $l$-th tuple, we estimate $\psi_l : \mathbb{R}^L \to \mathbb{R}$ using Gaussian process regression to minimize the following value.

$$\sum_{i=1}^{m-1} \left| x_k(t_{i+1}) - \psi_l\left(x_{l_1}(t_i), x_{l_2}(t_i), \cdots, x_{l_L}(t_i)\right) \right| \tag{2}$$

After that, we estimate the probability density function $p(x)$ by performing kernel density estimation from the set of estimates obtained by calculating one step estimation $\tilde{x}_l^k(t+1) = \psi_k^l\left(x_{l_1}(t), \cdots, x_{l_L}(t)\right)$ from each $\psi_l$.

And we calculate the skewness $\gamma$ of the probability density function, and if $\gamma$ is 0.5 or less, it is adopted and $\tilde{x}_k(t+\tau) = \int xp(x)\mathrm{d}x$ is determined as estimation. If not, we correct the estimate as follows.

We calculate the in-sample error $\delta_l (>0)$ and pick the $\left[\frac{n}{2}\right]$ best samples and estimate accordingly.

$$\tilde{x}_k(t+\tau) = \sum_{i=1}^{\left[\frac{n}{2}\right]} \omega_i \tilde{x}_k^{l_i}(t+\tau), \quad \omega_i = \frac{\exp\left(-\frac{\delta_i}{\delta_1}\right)}{\sum_j \exp\left(-\frac{\delta_j}{\delta_1}\right)} \tag{3}$$

In this work, the estimation period is 2018 and the estimation is performed with $L=10$ and $s=3$. The data is the intraday returns of each stocks included in each industry. Then, we predict the intraday returns of each stocks in each industry one period each, and calculate the average value of the mean squared error (MSE) in the whole industry from the actual intraday return over the entire prediction period is an index for prediction accuracy.

As a comparison target, we calculate the average value of MSE with the actual return when each stock is predicted by simple linear regression and LASSO regression when $L=10$ using other stocks of the industry without using the randomly distributed embedding method.

### 3.2. Verification Result

We show the result of the experiment in Table 1. As a result, the random distribution embedding method became the most accurate method in all industries. Compared with the other industries, the scope of improvement of this method is larger in food and electronics.

Table 1. Results.

| | Linear | LASSO | RDE |
|---|---|---|---|
| Construction | 0.95436 | 0.51920 | 0.46580 |
| Chemistry | 3.31570 | 3.19647 | 2.77994 |
| Food | 1.25392 | 0.68217 | 0.44245 |
| Machinery | 2.98313 | 2.23938 | 1.63249 |
| Electronics | 3.17653 | 2.67326 | 1.65308 |
| Pharmaceuticals | 1.81477 | 1.31412 | 0.98475 |
| Transportation | 1.05024 | 0.79956 | 0.74680 |

As a premise, in order for the randomly distributed embedding method to work, the variables to be analyzed must be in the same attractor. In that sense, compared to the other industries, we can guess that the stocks included in the food and electronics industry are on the same attractor, that is, the relationship between the stocks is relatively close.

## 4. Conclusions

In this work, we showed that we could improve the prediction accuracy when we use the randomly distributed embedding method, which is the method of randomly selecting variables from the values of many observational variables at a certain time and estimating the attractor state at that time, for predicting future returns of Japanese stocks comparing with the time when we use simple linear regression or LASSO regression. In addition, it can be inferred that the improvement range of the prediction accuracy is different depending on the type of industry, the nature of the stock group included in the type of industry and the degree to which these stocks are in the same attractor.

As a future perspective of this work, it is possible to aim for more accurate forecasting accuracy by applying randomly distributed embedding method to financial instruments that are likely to be on the same attractor, such as multiple volatility indexes. In addition, it is possible to aim to improve the prediction accuracy by using an algorithm such as LSTM as a regression method used for the randomly distributed embedding method. Furthermore, it is possible to use, for example, for stock selection filtering in investment methods in which the closeness of the nature between stocks is important, such as pair trade, by using the prediction accuracy improvement range from the conventional method according to the randomly distributed embedding method.

## Acknowledgements

This paper does not represent official views of Epic Partners Investments Co., Ltd. and the University of Tokyo to which the authors belong. Everything is the personal opinion.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Ma, H., Leng, S., Aihara, K., Lin, W. and Chen, L. (2018) Randomly Distributed Embedding Making Short-Term High-Dimensional Data Predictable. *PNAS*, **115**, E9994-E10002. https://doi.org/10.1073/pnas.1802987115

[2] Sauer, T., Yorke, J.A. and Casdagli, M. (1991) Embedology. *Journal of Statistical Physics*, **65**, 579-616. https://doi.org/10.1007/bf01053745

[3] Takens, F. (1981) Detecting Strange Attractors in Turbulence. In: Rand, D.A. and Young, L.-S., Eds., *Dynamical Systems and Turbulence*, Springer, Berlin, 366-381.

https://doi.org/10.1007/bfb0091924

[4] Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer, Berlin, 303-319.

[5] Deyle, E.R. and Sugihara, G. (2011) Generalized Theorems for Nonlinear State Space Reconstruction. *PLoS One*, **6**, e18295. https://doi.org/10.1371/journal.pone.0018295