Index Fund Selections with GAs and Classifications Based on Turnover

Yukiko Orito* Member
Takaaki Motoyama** Non-member
Genji Yamazaki*** Non-member

It is well known that index fund selections are important for the risk hedge of investment in a stock market. The ‘selection’ means that for ‘stock index futures’, $n$ companies of all ones in the market are selected. For index fund selections, Orito et al. proposed a method consisting of the following two steps: Step 1 is to select $N$ companies in the market with a heuristic rule based on the coefficient of determination between the return rate of each company in the market and the increasing rate of the stock price index. Step 2 is to construct a group of $n$ companies by applying genetic algorithms to the set of $N$ companies. We note that the rule of Step 1 is not unique. The accuracy of the results using their method depends on the length of time data (price data) in the experiments. The main purpose of this paper is to introduce a more ‘effective rule’ for Step 1. The rule is based on turnover. The method consisting of Step 1 based on turnover and Step 2 is examined with numerical experiments for the 1st Section of Tokyo Stock Exchange. The results show that with our method, it is possible to construct a more effective index fund than the results of Orito et al. The accuracy of the results using our method depends little on the length of time data (turnover data). The method especially works well when the increasing rate of the stock price index over a period can be viewed as a linear time series data.

**Keywords:** Index Fund, Turnover, Genetic Algorithms, Tokyo Stock Exchange

1. Introduction

‘Stock index futures’ have been used very extensively for the hedge trading which is the practice of offsetting the price risk in any stock market position by taking an equal but opposite position in the futures market. Suppose that we select $n$ listed companies in a stock market and invest in $c$ stocks of each of the $n$ companies. Let $P(t)$ be the total price of the $n$ companies at time $t$, by assuming that the stock index future starts at $t = 0$. It is well known that the $n$ companies are very useful for the risk hedge if the total return rate, defined as $(P(t) − P(t − 1))/P(t − 1)$, follows well the increasing rate of the stock price index in the market (for this, see, e.g., (1),(4)). Such a group consisting of $n$ companies is called ‘index fund.’ Then we have an important problem of finding $n$ companies in the market whose total return rate follows well the increasing rate of the stock price index. In this paper, the problem is called ‘an index fund selection.’ The coefficient of determination between the total return rate and the increasing rate of the stock price index, $R^2$ (defined in Section 2), is usually used as a measure of how the return rate follows the increasing rate of the stock price index (see, e.g., (1)). Roughly speaking, the $R^2$ can be interpreted as the correlation coefficient between the return rate and the increasing rate of the stock price index. The index fund selection can be viewed as a combinatorial optimization problem, which means that it is to select the $n$ companies from the stock market such that the $R^2$ is high.

The index fund requires rebalancing in order to reflect the changes in the composition of the stock price index. However, the price of the stock is unknown, so the implied cost of rebalancing is uncertain. To construct the group consisting of $n$ companies, we have to make a great of investment in rebalancing when the number of companies in the group is large. From a practical viewpoint, hence, it is desired that the group consisting of $n$ companies is constructed such that $n$ is small but the $R^2$ is high. In this context, Takabayashi reported that for the 1st Section of Tokyo Stock Exchange (TSE) consisting of about 1400 companies, it was possible to construct a group with $n ≈ 440$ whose $R^2$ was about 0.96. He selected 900 companies in the market based on ‘the weighted mean’ and then applied Genetic Algorithms (GAs) to the 900 companies for the index fund selections. Orito et al. proposed a method consisting of the following two steps.

Step 1 Each company in the stock market is weighted (or assigned a value) according to a ‘heuristic rule.’ The set of $N$ companies is selected based on the values.
Step 2  The group consisting of \( n \) companies is constructed for the index fund by applying GAs to the \( N \) companies. Their heuristic rule is based on the coefficient of determination between the return rate of each company in the market and the increasing rate of the stock price index. They applied the method consisting of Step 1 based on the above rule and Step 2 to the 1st and 2nd Sections of TSE. They concluded from the case studies that it was possible to construct an effective group for the index fund with \( N = 200 \sim 300 \), not depending on the market and the period. For example, they constructed the group consisting of 92 (\( n \)) companies whose \( R^2 \) was about 0.96 over the above period of the 1st Section of TSE by setting \( N = 200 \). However, the accuracy of the results using their method depends on the length of time series data (price data) in the experiments. We note that the heuristic rule of Step 1 is not unique. The main purpose of this paper is to introduce a more ‘effective rule’ of Step 1. The rule is based on turnover.

To demonstrate the effectiveness of our method consisting of Step 1 based on the proposed rule and Step 2, we have applied our method to the 1st Section of TSE. The results show that with our method, it is possible to construct the group consisting of \( n \) companies in which \( R^2 \) is higher than the results of Orito et al.\(^{(6)} \). The accuracy of the results using our method depends little on the length of time data (turnover data).

2. Preliminaries

Suppose that a stock market consists of \( K \) companies, numbered company 1, company 2, \ldots, company \( K \). Suppose that we invest in a group consisting of \( n \) companies in a stock market consisting of \( K \) companies which starts at \( t = 0 \) and ends at \( t = T \). The ‘\( t \)’ is on date basis. It is assumed that throughout the paper, we invest \( n \) stocks of each company belonging to the group. This means that the portfolio is unique for the group.

In the field of regression analysis, the coefficient of determination (\( CD \)) has often used as a measure of how well an estimated regression fits. As the coefficient approaches 1, the estimated regression fits better (for this, see, e.g.,\(^{(3)} \)). By analogy with this, the \( CD \) between the return rate of the index fund and the increasing rate of the stock price index over the interval \([1, T]\), defined as

\[
R^2 = \frac{(TZ - XY)^2}{(TU - X^2)(TW - Y^2)},
\]

\[
X = \sum_{t=0}^{T} x(t), \quad Y = \sum_{t=0}^{T} y(t), \quad U = \sum_{t=0}^{T} x^2(t),
\]

\[
W = \sum_{t=0}^{T} y^2(t), \quad Z = \sum_{t=0}^{T} x(t)y(t)
\]

has been used as the fitness measure between the index fund and the stock price index. Here \( y(t) \) is the return rate of the index fund between time \( t - 1 \) and \( t \) and \( x(t) \) is the increasing rate of the stock price index between time \( t - 1 \) and \( t \). In this paper, the \( R^2 \) is adopted as the fitness measure.

3. Method for Index Fund Selections

As mentioned in Section 1, the method for index fund selections consists of the following two steps.

3.1 Step 1  The first step is to select \( N \) companies of \( K \) ones according to a heuristic rule. The rule of Orito et al.\(^{(6)} \)'s is based on the \( CD \) between the return rate of each company in the market and the increasing rate of the stock price index. On the other hand, we introduce the rule based on ‘turnover’ in this paper. For a company, (say) company \( i \) in the market, let

\[
V_i(T_0) = \frac{1}{T - (T_0 - 1)} \sum_{t=T_0}^{T} w_i(t)
\]

\[
(T_0 = 0, 1, \ldots, T), \quad \ldots \ldots (1)
\]

where \( w_i(t) \) is the turnover of company \( i \) at \( t = T_0 \) \((\leq T)\). Suppose that the \( V_i(T_0)'s \) are assigned to all companies in the market. Without loss of generality, we can renumber the all companies so that

\[
V_1(T_0) \geq V_2(T_0) \geq \cdots \geq V_i(T_0) \geq \cdots \geq V_K(T_0).
\]

We note that the renumbered company \( i \) has the \( i \)-th high \( V_i(T_0) \) in the all companies. Then the heuristic rule means that the set of company 1, company 2, \ldots, company \( N \) is selected. This is first step and the set of \( N \) companies is denoted by \( S_N(T_0) = \{ \text{company 1, company 2, \ldots, company } N \} \) or \( \{ 1, 2, \ldots, N \} \) if no confusion.

3.2 Step 2  The second (final) step is formulated as the problem of finding the group consisting of \( n \) companies, denoted by \( I_{n,N} \), such that the \( R^2 \) is the highest one in \( R^2 \)s of all subsets of \( S_N(T_0) \). It is well known that GAs are useful for such optimization problems (for this, see, e.g.,\(^{(3)} \)). Hence, we use a GA for constructing the group \( I_{n,N} \) from \( S_N(T_0) \). Suppose the \( S_N(T_0) \) is given. The approach of implementation of GA follows that of Orito et al.\(^{(6)} \). A gene is defined by

\[
g_i = \begin{cases} 
0 & \text{Company } i \text{ is not an element of } I_{n,N} \\
1 & \text{Company } i \text{ is an element of } I_{n,N} 
\end{cases} \quad (i = 1, 2, \ldots, N)
\]

and a chromosome is denoted by \( \bar{g} = \{ g_1, g_2, \cdots, g_N \} \). The fitness value of the GA is \( R^2 \). The GA is designed as follows:

(1) Start

We randomly generate 100 chromosomes as the initial population.

(2) Crossover

The crossover exchanges the partial structure between two chromosomes at two points selected at random, with the crossover rate \( P_c \).

(3) Mutation

The mutation replaces the partial structure between two points of one chromosome selected at random, 0 to 1 (or 1 to 0), with the mutation rate \( P_m \).
(4) Selection
Suppose that $M$ chromosomes, numbered $g_1$ through $g_M$, are selected after applying the crossover and the mutation. Let $f_i$ be the $R^2$ for $g_i$ and let $p_i = f_i / \sum_{j=1}^{M} f_j$. The (cumulative) probability $q_i = \sum_{j=1}^{i} p_j$ is assigned to $g_i$. We generate a random number $r$ in $[0, 1]$. The $g_i$ is selected as an element of the new population to the next generation when $q_{i-1} < r \leq q_i$. Repeating this procedure, we select 100 chromosomes. If the highest $R^2$ in the 100 chromosomes is less than that for $M$ chromosomes, say the $R^2$ for $g_{i_n}$, one of the 100 chromosomes is randomly replaced with $g_{i_n}$. The set of 100 chromosomes obtained by the procedure becomes the new population to the next generation.

(5) Stop
The GA is broken off on the 100th generation.
If the number of generations is less than 100, the GA goes back to the crossover.
On the 100th generation, the chromosome with the highest fitness value gives $I_{n,N}$ by the GA. For fixed $S_N(T_0)$, the GAs are executed 20 times and 20 groups are given. The average of the $R^2$s for 20 $I_{n,N}$'s is below referred to as the CDPA ('P' for portfolio and 'A' for average) of the $S_N(T_0)$. In this paper, the method proposed by Orito et al. (6) and our method are called 'Method 1' and 'Method 2', respectively.

4. Numerical Experiments
We have applied Method 2 to the 1st Section of TSE. In this section, the results for the following cases are shown.

- Data 1 (the 1st Section of TSE, $K = 1323$)
  - Period: Aug. 12, 1997 - Aug. 17, 1998 ($T = 250$)
- Data 2 (the 1st Section of TSE, $K = 1356$)
  - Period: Jan. 1, 1999 - Dec. 31, 1999 ($T = 245$)

The data are the same as those used by Orito et al. (6). As the stock price index in Data 1 or 2, the Tokyo Stock Price Index (TOPIX) is used and the movement of TOPIX normalized with the value at $t = 0$ is shown by a thin line in Fig. 1 (a) and (b) or Fig. 2 (a) and (b), respectively.

4.1 Preliminary Experiments
We have performed the preliminary experiments in order to set the GA parameters, the crossover rate $P_c$ and the mutation rate $P_m$. The pairs $(P_c, P_m)$ used in the experiments are all combinations of $P_c = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ or $0.9$ and $P_m = 0.025, 0.05, 0.075, 0.1, 0.125, 0.15,$ or $0.2$. For Data 1, we picked up several sets consisting of $S_N(T_0)$. Each set consists of 100, 200, 300 or 400 companies in the case of $T_0 = T$. For each set, the CDPA has been obtained for the pairs $(P_c, P_m)$. The results say that we had better use the pair $(P_c, P_m) = (0.9, 0.05)$ or $(P_c, P_m) = (0.9, 0.75)$. Hence the $(P_c, P_m)$ in below experiments is set as $(0.9, 0.05)$.

4.2 Main Results
For Data 1 or 2, we had carried out the numerical experiments of Method 2 as $N = 200$. The group $I_{n,200}$ can be constructed by applying Step 2 to $S_{200}(T_0)$ based on Eq. (1) in Data 1 or 2. For each data, the results of the GA 20 times for $S_{200}(T_0)$ in the case of $T_0 = 0$ is shown in Table 1 respectively and the CDPA is shown by a thick line in Fig. 3 (a) or (b) as a function of $T_0$. Similarly, we can construct the set of $N$ companies which is based on the CD between the return rate of company $i$’s stock and the increasing rate of the stock price index over $[T_0, T]$ (Method 1). The CDPA obtained by the groups constructed with Method 1 is shown by a thin line in Fig. 3 as a function of $T_0$. 

Fig. 1. The return rate of $I_{n,200}$ and the increasing rate of the stock price index in Data 1 as a function of $t$

Fig. 2. The return rate of $I_{n,200}$ and the increasing rate of the stock price index in Data 2 as a function of $t$
Table 1. The results obtained by the GA 20 times for $S_{200}(0)$

| No. | Data 1   | | Data 2   | |
|-----|----------||----------||
|     | CDP      | n  | CDP      | n  |
| 1   | 0.9786   | 100| 0.8959   | 117|
| 2   | 0.9781   | 113| 0.8966   | 108|
| 3   | 0.9810   | 108| 0.9003   | 119|
| 4   | 0.9811   | 116| 0.9053   | 117|
| 5   | 0.9809   | 103| 0.9029   | 116|
| 6   | 0.9801   | 113| 0.9035   | 105|
| 7   | 0.9811   | 91 | 0.9027   | 114|
| 8   | 0.9776   | 107| 0.9045   | 118|
| 9   | 0.9811   | 125| 0.9025   | 117|
| 10  | 0.9785   | 102| 0.8886   | 115|
| 11  | 0.9784   | 105| 0.9030   | 116|
| 12  | 0.9802   | 109| 0.8927   | 117|
| 13  | 0.9813   | 106| 0.8951   | 115|
| 14  | 0.9774   | 102| 0.8991   | 120|
| 15  | 0.9810   | 103| 0.8975   | 114|
| 16  | 0.9822   | 116| 0.8943   | 100|
| 17  | 0.9775   | 112| 0.9024   | 116|
| 18  | 0.9814   | 107| 0.9015   | 104|
| 19  | 0.9776   | 122| 0.9054   | 120|
| 20  | 0.9831   | 95 | 0.8957   | 103|

CDPA  0.9798  0.8995
Max. CDP  0.9831  0.9054
Min. CDP  0.9774  0.8886
Range of n  91~125  100~120

Fig. 3. The CDPA obtained by Method 1 or 2 as a function of $T_0$

For Data 1, Orito et al. (6) constructed the group consisting of 82 ~ 104 companies with Method 1 as $N = 200$ whose CDPA was 0.9574. Comparing this with the results in Table 1, we can conclude that it is possible to construct the group with Method 2 whose $R^2$ is higher than that of Method 1. For Data 2, we have similar results to those for Data 1.

Fig. 3 suggests that the CDPA of the group constructed by Method 2 with $S_{200}(T_0)$ depends little on the $T_0$ and it is higher than that of Method 1. This means that Method 2 gives the effective index fund $I_{n,200}$. The accuracy of the results using Method 2 depends little on the length of time series data (turnover data). Thus, Method 2 consisting of Step 1 based on turnover and Step 2 is effective for index fund selections.

4.3 The Difference of the CDPA between Data 1 and 2 In this paper, we use the $R^2$ as the measure of how the return rate of $I_{n,200}$ follows the stock price index, i.e., of the fitness. Our numerical experiments show that for a fixed $S_n(T_0)$, the $R^2$'s obtained by Method 2 in Data 1 are higher than the results of Data 2. What is the main difference between Data 1 and 2? In this section, we discuss this problem.

When the increasing rate of the stock price index over a period is viewed as a time series data, the data can be characterized as (i) a linear model over the period, or (ii) a non-linear (or chaotic) model which means that there are non-linear time intervals in the period (see, for example, Tanaka et al. (5)). The fitness measure, $R^2$ is selected based on (i). False Nearest Neighbors (FNN) analysis is a useful method to examine if the stock price index data over a period can be viewed as (i) (5). The FNN analysis is a kind of dimension estimation methods for time series data and is summarized as follows. The distances on $m$- and $(m + 1)$- dimensional phase spaces for a time series data over a period, $D(m)$ and $D(m + 1)$, are calculated. Using $D(m)$, $D(m + 1)$ and a fixed evaluation criterion $d_0 (0 < d_0 < 1)$, the false rate, $d(m)$, between the $m$- and $(m + 1)$- spaces for the data is calculated. By the movement of $d(m)$ as a function of $m$, we judge if the data can be viewed as (i).

We have applied the FNN analysis to Data 1 and 2 as $d_0 = 0.3$. The results for Data 1 or 2 is illustrated in Fig. 4 (a) or (b), respectively.

As shown (a) in Fig. 4, the $d(m)$ of the stock price index for Data 1 is a decreasing function of $m$. Then the stock price index data can be viewed as (i). For this case, it is possible to construct the index fund selections whose $R^2$'s are very high. On the other hand, Fig. 4 (b) for Data 2 shows that the $d(m)$ first decreases, increases, and decreases again as $m$ increases. That is, the $d(m)$ as the function of $m$ forms a valley. Then, the stock price index data can not be viewed as (i). To obtain higher fitness for (ii), we may need further work such as
introducing other measures of the fitness.

For Data 1 or 2, the CDPA obtained by Method 2 is 0.9798 which is close to the $R^2$ of #6 $I_{113,200}$ or is 0.8995 which is close to the $R^2$ of #14 $I_{120,200}$ in the table, respectively. The movement of the total return rate of $I_{113,200}$ or $I_{120,200}$ given normalized total return rate at $t = 0$ is shown by a thick line in Fig. 1 (a) or Fig. 2 (a), respectively. For the CDPA obtained by Method 1 in Data 1, the movement of the total return rate of $I_{92,200}$ (which is close to the CDPA) given normalized total return rate at $t = 0$ is also shown by a thick line in Fig. 1 (b). Similarly, the movement obtained by Method 1 in Data 2 is shown in Fig. 2 (b). As expected, the fitness between the index fund and the stock price index in Data 1 is very well.

5. Concluding Remarks

This paper proposed a simple method for constructing the index fund, $I_{n,N}$, whose return rate followed well the increasing rate of the stock price index in the stock market. We have applied the method to the 1st Section of TSE.

The numerical experiments showed that our method gave the effective index fund whose the fitness was higher than the results of Orito et al. (6). The accuracy of the results using our method depends little on the length of time series data (turnover data). The method especially works well when the increasing rate of the stock price index over a period can be viewed as a linear time series data. The effectiveness of our method in forecasting future stock market remains to be seem. This will be revealed in future studies.

Acknowledgment

We are grateful to the referees whose valuable comments helped to improve the content of this paper.

(Manuscript received Dec. 11, 2003, revised June 11, 2004)

References


Yukiko Orito (Member) was born in Tokyo, Japan, on July 11, 1974. She received a Ph.D. degree in systems engineering science from Tokyo Metropolitan Institute of Technology in 2003, and is presently a lecturer at Ashikaga Institute of Technology.

Takaaki Motoyama (Non-member) was born in Tokyo, Japan, on June 10, 1978. He received a master’s degree in intelligent systems from Tokyo Metropolitan Institute of Technology in 2003, and is presently a systems engineer at Canon Inc.

Genji Yamazaki (Non-member) was born in Hokkaido, Japan, on October 8, 1946. He received a Ph.D. degree in mechanical engineering from Kogakuin University in 1978, and is presently a professor at Tokyo Metropolitan Institute of Technology.