Metaheuristics for Sparse Index-Tracking Problem: A Case Study on FTSE 100

Man-Chung Yuen, Sin-Chun Ng, and Man-Fai Leung*

The Open University of Hong Kong, School of Science and Technology, Hong Kong, China

*Email: mflung@ouhk.edu.hk

Abstract. Passive management contributes a more stable return than an active management strategy over the long term. Index-tracking is one of the passive investment strategies that attempt to replicate market indexes to reproduce the performance. Sparse index-tracking considers a subset of market index stocks to minimize the difference between the market index and the replicated index. In this paper, two metaheuristics are applied to solve this problem. The sparse index-tracking problem formed by the objective function of the empirical tracking error with the penalty values that result in an NP-hard problem. The penalty value is used to restrict the numbers of the considered stocks. To show the performance of the metaheuristics, various penalty values are investigated, and they produce approximation solutions to the index-tracking problem. Among them, particle swarm optimization shows better or statistically similar performance to GA in solving the sparse index-tracking problem.

1. Introduction

On the stock exchange, active and passive management are two main types of investments. The fund manager plays an essential role in the financial market for the management of portfolio trading equities based on the fund's investment strategy. To represent market performance in the financial world, the stock index is determined as a subset of financial assets, and is estimated using the values of certain securities by the weighted arithmetic average. Recently, the percentage of the passive fund has increased and is outperforming that of the active management funds [1, 2]. The fund manager seeks for long-term returns through the investment of assets such as [3] securities and equities. In a few decades, the fund management industry is a fast-growing business. Public is more interested in the diversification of portfolio investment risks. The concept of active fund management [4] and passive fund management is under discussion.

For an active management approach, the active fund manager does not believe in the efficient market hypothesis [5]. The manager thinks the market information is not perfectly efficient. The manager is possible to beat the market through his judgment, professional expertise, and experience for the decision making to trade the stocks such as market analysis, investment investigation, and price predictions. In general, the manager frequently buys and sells equities to take advantage of fluctuations in market prices, and has a higher level of flexibility to exchange equities.

On the contrary, the passive management approach believes in the efficient market hypothesis [6, 7]. This assumption assumes the asset, and stock price could fully reflect all marketing pieces of information at all times that implies the manager is not possible to circumvent the market [8]. Grounded on this hypothesis, the manager invests in mutual funds and exchange-traded funds (ETFs)
that portfolio's securities mirror the index fund such as S&P 500 and FTSE 100. In common, the
manager has less level of flexibility in the market, and he reaches the profitability similar to the
specified market index value with a closely determined set of criteria [9].
In the past few decades, the earnings of the passive management approach performed better than
the active management approach. Historical data show the growth of the stock index over the long
term. As a result, the total market capitalisation of ETFs has increased in different countries like
America and Europe [10]. Numerous researchers investigate the performance of ETFs [11]. In this
paper, particle swarm optimization (PSO) [12] and genetic algorithm (GA) [13] are used to solve the
index-tracking problem. The performance of the two metaheuristics is also presented in this paper.

2. Preliminaries and Problem Formulation

2.1. Passive Portfolio Management
The reason why passive management strategies outperform active management could know by their
characteristics [14-17]. (1) Passive management has a lower transaction cost than active management
because it sells the stocks rarely within the period. (2) Even professional investors do not always able
to acquire from the stock market. For instance, an investor may fail to invest in several years and only
beat the securities industry in a single year. (3) A proactive management strategy carries a higher risk
than passive management. The active management approach is exposed to both corporate and market
risk while passive management is exposed only to market risk. The stock market index also increased
continuity in the past. Consequently, positive returns are achieved on a long-term basis.
ETFs are categorized in full replication and sparse tracking index [18-20]. We are going to
investigate the latter one in this paper, and both advantages and disadvantages are presented. First, the
simplest way to deal with ETF is full replication that means the fund purchase all the stocks to
represent the stock index. Under this methodology, a perfect tracking error is achieved. While at the
same time, complete replication contains large quantities of shares that could produce a huge
transaction cost. In addition, it exists many small quantities of stocks that are inconvenient to sell them
and costly to buy them. Additionally, when larger assets under management, the full replication
performs better as the fixed cost no longer dominates the outcome. For the latter one, sparse index-
tracking tries to track the stock index with smaller amounts of assets to reproduce the stock market
index approximately [21-23]. When small amounts of assets, it implies fewer transaction costs because
of the minimum fixed commission cost reduced. Furthermore, when smaller assets are under
management, the monitoring of the sparse index gives better results, because fixed costs significantly
dominate the result.

There are two general methods to solve this problem. For the first method, this method
decomposes the whole task into a two-step procedure for stock selection and capital allocation. Several
approaches have been proposed for the stock selection step, such as the capitalization-
weighted grouping technique, heuristic procedures [24-27]. Once the subset of assets has been selected,
the next step determines the portfolio weights based on the selected stocks. While this method is
straightforward, its optimization is not clear. Besides, the selection process is a NP hard problem.
Another way to meet this challenge is to combine a two-step procedure into a single automation
process [22, 28-29]. The difficulties in selecting securities and weighting are explained in a single task
by the penalty value. The penalty value penalizes the cardinality constraint for the sparsity of stocks.

2.2. Problem Formulation
Metaheuristic methodologies applied to the index-tracking problem. PSO and GA are widely used
evolution algorithm. The populations explore in the feasible space to find considerably approximating
solutions. However, they are no guarantee that an optimal solution will be found.
The PSO algorithm was developed by Kennedy and Eberhart. It is based on swarm behaviour such
as bird flocking. Each particle updates its position according to its personal best position and the
global best position in the multidimensional searching space.
\[ x_{v+1} = x_v + w \cdot v + c1 \cdot (pbest - x_v) + c2 \cdot (gbest - x_v) \]  

The GA imitates the natural selection mechanism from Darwin’s theory. This algorithm is inspired by Holland. The GA simulates the biological behavior that evolution based on the rule of survival of the fittest value. It contains selection, crossover, and mutation. These metaheuristics applied to this problem that aims to minimize the between for returns of the portfolio funds and benchmark funds.

The lookback approach tries to look at the historical information to track the index in a specific period. The empirical tracking error defined as

\[ \text{ETE} = \frac{1}{T} || rp - rb ||_2^2 \]  

where \( rp = \sum_{t=1}^{T} \sum_{n=1}^{N} r_{t,n} \odot w_{t,n} \), \( T \) is maximum trading days, \( N \) is the total numbers of available stocks, \( rp \) is the portfolio returns in time \( t \), and \( rb \) is return of the benchmark index in time \( t \), and \( n \) is the current stock in the index-tracking problem. In addition, the \( \odot \) is the hadamard product. Note that the weights of the decision variable repeat copies at time \( t \). The objective function with the penalty value shown as follow [28].

\[ \text{ETE with penalty value} = \frac{1}{T} || rp - rb ||_2^2 + pe, \]  

where \( pe = \sum_{n=1}^{N} \lambda 1^T \frac{\log(1+|w_n|/p)}{\log(1+\gamma/p)} \), \( p \) is 10³ for good approximation [28], \( \gamma \) is the upper bound of the current weights, \( \lambda \) is the penalty value. Note that the least absolute shrinkage and selection operator (LASSO) [30] is not suitable for this equation because the summation of the weights is fixed to one for a constant value. The return of the index in time \( t \) calculated as follow.

\[ r = \frac{p_t - p_{t-1}}{p_{t-1}} \]  

3. Experimental Results

For simplicity, the dataset follows the standard Pareto principle of the well-known 80/20 rule [31]. This rule exposes the results of approximately 80% of the outcomes originating from 20% of the causes. Most machine learnings also split the dataset with this approach (e.g., [32]). Figure 1 shows the training dataset and the testing dataset are 8 months and 2 months, respectively.

![Figure 1. Training dataset and testing dataset.](image-url)

Furthermore, the close prices of the index and these companies of the daily return are used. However, the daily trading days are not available for every day. For more precise, the total trading days are equally divided into the numbers of months. Thus, it avoids different periods for the dataset. For simplicity, when the companies are no longer on the market index or not available in the period, these indexes and returns will be set to zero. The dataset period is from 01-01-2016 to 12-31-2016.

For a fair comparison, the stopping criteria of these algorithms are the fitness value that does not change in 100 iterations for each simulation. The population sizes are set to 1000 for both algorithms. The inertia weight is 0.7, and acceleration parameter \( c1 \) and \( c2 \) are 2.0 for PSO. For GA, the roulette wheel selection (RWS) and binary tournament selection (BTS) are tested. The mutation rate set for 0.02, the uniform crossover is used, and selection pressure is 1.0 for roulette wheel selection. Ten independent runs for the simulation. The results of the average iterations, ratio of zeros, average
training error, average testing error, best training error, and best testing error are presented. In the simulation result, PSO and GA show their outcomes for this problem. The performance of PSO is shown in Table 1. The performances of GA with RWS and BTS are shown in the Table 2, and Table 3, respectively. When the penalty value larger, PSO performs better than GA. PSO also obtains larger ratios of zero than GA in all test cases.

Table 1. Particle swarm optimization for index-tracking.

| Penalty value λ | Average iterations | Ratio of zeros | Mean training error | Mean testing error | Best training error | Best testing error |
|-----------------|-------------------|----------------|--------------------|--------------------|--------------------|--------------------|
| 1e-0            | 133               | 0.810          | 3.05E-03          | 2.12E-03          | 1.37E-03          | 1.36E-03          |
| 1e-2            | 132               | 0.806          | 3.23E-03          | 1.79E-03          | 1.36E-03          | 1.26E-03          |
| 1e-4            | 133               | 0.802          | 2.93E-03          | 2.04E-03          | 1.27E-03          | 1.05E-03          |
| 1e-6            | 201               | 0.771          | 6.88E-04          | 7.48E-04          | 4.92E-04          | 4.64E-04          |
| 1e-8            | 5535              | 0.523          | 7.67E-05          | 2.37E-04          | 5.08E-05          | 1.55E-04          |
| 1e-10           | 5816              | 0.452          | 7.19E-05          | 2.33E-04          | 5.23E-05          | 1.81E-04          |
| 0               | 4531              | 0.438          | 6.55E-05          | 2.24E-04          | 4.15E-05          | 1.89E-04          |

Table 2. Genetic algorithm with roulette wheel selection (RWS) for index-tracking.

| Penalty value λ | Average iterations | Ratio of zeros | Mean training error | Mean testing error | Best training error | Best testing error |
|-----------------|-------------------|----------------|--------------------|--------------------|--------------------|--------------------|
| 1e-0            | 1648              | 0.502          | 3.28E-02          | 3.16E-02          | 6.16E-03          | 8.50E-03          |
| 1e-2            | 1662              | 0.481          | 2.10E-02          | 1.71E-02          | 6.83E-03          | 9.34E-03          |
| 1e-4            | 1625              | 0.478          | 6.33E-03          | 8.22E-03          | 5.80E-03          | 6.58E-03          |
| 1e-6            | 3983              | 0.428          | 4.51E-04          | 8.60E-04          | 3.79E-04          | 5.67E-04          |
| 1e-8            | 1391              | 0.305          | 4.11E-05          | 1.80E-04          | 3.11E-05          | 1.66E-04          |
| 1e-10           | 1005              | 0.219          | 4.64E-05          | 1.80E-04          | 3.23E-05          | 1.51E-04          |
| 0               | 1026              | 0.207          | 4.58E-05          | 1.73E-04          | 3.06E-05          | 1.38E-04          |

Table 3. Genetic algorithm with binary tournament selection (BTS) for index-tracking.

| Penalty value λ | Average iterations | Ratio of zeros | Mean training error | Mean testing error | Best training error | Best testing error |
|-----------------|-------------------|----------------|--------------------|--------------------|--------------------|--------------------|
| 1e-0            | 1054              | 0.501          | 3.10E-02          | 2.30E-02          | 6.66E-03          | 1.03E-02          |
| 1e-2            | 1047              | 0.479          | 2.87E-02          | 2.05E-02          | 5.80E-03          | 1.10E-02          |
| 1e-4            | 1029              | 0.461          | 6.60E-03          | 1.01E-02          | 5.80E-03          | 7.58E-03          |
| 1e-6            | 2388              | 0.413          | 4.39E-04          | 8.01E-04          | 3.51E-04          | 5.83E-04          |
| 1e-8            | 2711              | 0.381          | 3.26E-05          | 1.89E-04          | 2.78E-05          | 1.64E-04          |
| 1e-10           | 2773              | 0.231          | 2.90E-05          | 1.65E-04          | 2.49E-05          | 1.50E-04          |
| 0               | 2586              | 0.218          | 2.89E-05          | 1.63E-04          | 2.46E-05          | 1.55E-04          |
On another side, GA performs better than PSO when the penalty value smaller. However, when the penalty value smaller, PSO performs slightly worse than GA. In GA, BTS presents slightly better than RWS in most test cases. GA requires more iterations than PSO on average. Additionally, both algorithms suffer the difficulty of premature convergence when the penalty value increases. The sparsity of this problem brings complexity to these original PSO and GA.

Moreover, we presented the total change in investment over the period by the cumulative return. Note that cumulative return is calculated by cumulative product for the return. Due to the page limit, only some parts of the cumulative return are shown in figures 2 and 3.

In this paper, the original PSO and GA are applied for the index-tracking problem. Numerous enhanced PSO and GA are proposed to improve the ability of exploration, exploitation. The Improved PSO, Adaptive PSO, and competitive swarm optimizer (CSO) are proposed for improving the inertia weight, acceleration parameters, and new learning strategy [33-37]. CSO was also suitable for large-scale optimization that may benefit from huge stock tracking. Various selection operators, crossover operators, and adjacency mutation operators for enhancing GA [38, 39]. When the penalty value increases, the complexity will also increase. These algorithms may obtain better accuracy and avoid dropping into local minima for this problem by these enhanced methodologies.

4. Conclusion
In conclusion, different metaheuristics are used to solve the sparse index-tracking problem. Generally, PSO performs better than the genetic algorithm in the simulation results. In particular, performs better than GA in terms of training error with larger penalty values. The average iterations, the ratios of zero, training error and testing error are measured for the sparse index-tracking problem. However, the high dimensional data and high penalty value require a large number of iterations as algorithms may stuck at the local minimum.
Figure 3. Cumulative return with penalty value 1e-8.

References
[1] Malkiel B G 2003 Passive investment strategies and efficient markets. European Financial Management 9(1) 1-10
[2] Anadu K, Krutli M S, McCabe P E, Osambela E, and Shin C 2019 The shift from active to passive investing: potential risks to financial stability? Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3244467
[3] Walden M L 2015 Active versus passive investment management of state pension plans: implications for personal finance Journal of Financial Counseling and Planning 26(2) 160-171
[4] Leung M F and Wang J 2020 Minimax and biobjective portfolio selection based on collaborative neurodynamic optimization. IEEE Transactions on Neural Networks and Learning Systems. In press. doi: 10.1109/TNNLS.2019.2957105
[5] Foster F D and Warren G J 2016 Interviews with institutional investors: The how and why of active investing. Journal of Behavioral Finance 17(1) 60-84
[6] Rudd A 1980 Optimal selection of passive portfolios. Financial Management 9(1) 57-66
[7] Stint B and Lewis J 1992 Guidelines for Rebalancing Passive-Investment Portfolios. Journal of Financial Planning, 5(2) 80
[8] Fama E F 1970 Efficient capital markets: A review of theory and empirical work. The Journal of Finance 25(2) 383-417
[9] Fraš A and Rogowski W 2016 The attractiveness of passive forms of investment in Poland. Journal of Management and Financial Sciences 25(3) 43-60
[10] Blitz D, Huij J, and Swinkels L 2012 The performance of European index funds and exchange traded funds. European Financial Management 18(4) 649-662
[11] Petruno T 2017 Small investors’ move to ‘passive’ stock funds becomes a stampede (Los Angeles Times)
[12] Kennedy J, and Eberhart R 1995 Particle swarm optimization. In Proceedings of ICNN’95-International Conference on Neural Networks (IEEE) vol 4 pp 1942-1948
[13] Holland J H 1992 Genetic algorithms. Scientific American 267(1) 66-73
[14] Woolley P, and Bird R 2003 Economic implications of passive investing. Journal of Asset Management 3(4) 303-312
[15] Sorensen E H, Miller K L, and Samak V 1998 Allocating between active and passive management. Financial Analysts Journal 54(5) 18-31
[16] Kremitzter K 2012 Comparing Active and Passive Fund Management in Emerging Markets. (University of California, Berkeley, Economics Department) Senior Honors Thesis
[17] Sushko V and Turner G 2018 The implications of passive investing for securities markets. BIS Quarterly Review, March 2018
[18] Pope P F and Yadav P K 1994 Discovering errors in tracking error. *Journal of Portfolio Management* 20(2) 27
[19] Frino A, Gallagher D R, Neubert A S, and Oetomo T N 2004 Index design and implications for index tracking. *The Journal of Portfolio Management* 30(2) 89-95
[20] Chen C, and Kwon R H 2012 Robust portfolio selection for index tracking. *Computers & Operations Research* 39(4) 829-837
[21] Beasley J E, Meade N, and Chang T J 2003 An evolutionary heuristic for the index tracking problem *European Journal of Operational Research* 148(3) 621-643
[22] Jansen R and Van Dijk R 2002 Optimal benchmark tracking with small portfolios. *The Journal of Portfolio Management* 28(2) 33-94
[23] García F, Guijarro F, and Oliver J 2018 Index tracking optimization with cardinality constraint: a performance comparison of genetic algorithms and tabu search heuristics. *Neural Computing and Applications* 30(8) 2625-41
[24] Oh K J, Kim T Y, and Min S 2005 Using genetic algorithm to support portfolio optimization for index fund management. *Expert Systems with Applications* 28(2) 371-9
[25] Li Y, Yang X, Zhu S, and Li D H 2014 A hybrid approach for index tracking with practical constraints. *Journal of Industrial & Management Optimization* 10(3) 905
[26] Dose C and Cincotti S 2005 Clustering of financial time series with application to index and enhanced index tracking portfolio. *Physica A: Statistical Mechanics and its Applications* 355(1) 145-151
[27] Bianchi D and Gargano A 2011 High-dimensional index tracking with cointegrated assets using an hybrid genetic algorithm. Manuscript available at: http://ssrn. com/abstract, 1785908
[28] Benidis K, Feng Y, and Palomar D P 2017 Sparse portfolios for high-dimensional financial index tracking *IEEE Transactions on Signal Processing* 66(1) 155-170
[29] Feng Y and Palomar D P 2016 A Signal Processing Perspective on Financial Engineering (Now Publishers) Vol. 9
[30] Tibshirani R 1996 Regression shrinkage and selection via the lasso *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1) 267-288
[31] Nisonger T E 2008 The “80/20 rule” and core journals. *The Serials Librarian* 55(1-2) 62-84
[32] Islam M M, Tayan O, Islam M R, Islam M S, Nooruddin S, Kabir M N, and Islam M R 2020 Deep Learning Based Systems Developed for Fall Detection: A Review. *IEEE Access* 8 166111-37
[33] Liu B, Wang L, Jin Y H, Tang F, and Huang D X 2005 Improved particle swarm optimization combined with chaos. *Chaos, Solitons & Fractals* 25(5) 1261-71
[34] Zhan Z H, Zhang J, Li Y, and Chung H S H 2009 Adaptive particle swarm optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39(6) 1362-81
[35] Cheng R, and Jin Y 2014 A competitive swarm optimizer for large scale optimization. *IEEE Transactions on Cybernetics* 45(2) 191-204
[36] Yuen, M C, Ng, S C, & Leung, M F 2020 A competitive mechanism multi-objective particle swarm optimization algorithm and its application to signalized traffic problem. Cybernetics and Systems, In press, DOI:10.1080/01969722.2020.1827795
[37] Leung, M F, Coello, C A C, Cheung, C C, Ng, S C, and Lui, A K F 2020 A hybrid leader selection strategy for many-objective particle swarm optimization. *IEEE Access* 8 189527-189545
[38] Leung F H F, Lam H K, Ling S H, and Tam P K S 2003 Tuning of the structure and parameters of a neural network using an improved genetic algorithm. *IEEE Transactions on Neural networks* 14(1) 79-88
[39] Dandy G C, Simpson A R, and Murphy L J 1996 An improved genetic algorithm for pipe network optimization. *Water Resources Research* 32(2) 449-458