J. Appl. Environ. Biol. Sci., 5(12S)116-121, 2015

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

Portfolio Selection for Index Tracking Using Independent Component Analysis

Bahareh Dadgar[1], Seyed Milad Rezvanizaniani[1], Pejman Mehran[2]

[1]Master of Science in Financial Engineering, Amirkabir University (Tehran Polytechnique), Iran
[3]Assistant Professor, Amirkabir University (Tehran Polytechnique), Iran

Received: July24, 2015
Accepted: September 31, 2015

ABSTRACT

The main issue in index tracking problem is to constructing a portfolio which aims at replicating the performance of a market. In this research, we used an integer model to construct the tracking portfolio by two similarity measures. One of them is Correlation and the other one is Information Coupling based on Independent Component Analysis (ICA). In this study, we demonstrate utility of our model for tracking NASDAQ100 index. At last, we show thatusing independent component analysis reduces the error of the tracking portfolio in comparison with using correlation.
KEYWORDS: Index Tracking, Integer Programming, Independent Component Analysis, Correlation

1. INTRODUCTION

Before 1952, investors used to focus on selecting individual stocks. So those who had less financial affordability would buy safer stocks and riskier stocks was purchased by rich investors who could tolerate more fluctuation. There were no measurement for portfolios risks until Harry Markowitz brought a new concept. He said that the portfolio risk is not the risk of each investment but the whole portfolio have a risk which is different from the individual investments risks [1]. So the best portfolio in a specific level of risk is the one that has the highest expected return. Based on the Sharp's Capital Asset Pricing model, in the market risk level no other portfolios can beat the market portfolio. He believes that if such a portfolio exist, the law of supply and demand put its security prices back in the line [2].

Investment managers invest million dollars per year and their aim is to benefit from both capital growth and dividend yield. They usually follow one of the following strategies [3]:

- Active investment, in this strategy fund managers have considerable latitude in their work. It is expected that they can pick outperforming stocks based on their knowledge and experiences. They usually try to gain higher return from their portfolios by admitting more risk (Speculation).
- Passive investment, in this strategy fund managers have less degree of flexibility and usually they have a specific objective: to establish a portfolio in order to gain the market return. So they usually track the main market index.
- Hybrid investment, above strategies are the main strategies of investing, it is possible to make a portfolio half actively and half passively.

Both active investment and passive investment strategies have their pros and cons:

- Active investment needs an expert management team that needs high payments which force high fixed cost to the fund. Moreover, in active investment due to speculation and frequent trading, the transaction cost is so high. In this situation, investor face both systematic risk (market risk) and unsystematic risk (Company- or industry-specific hazard that is inherent in each investment).
- Passive investment has less fixed cost because there is no need to high qualified experts for the fund management team. Furthermore the management team usually buy stocks and hold them for a middle to long time so transaction costs are much less than active investment. However, the main weakness of this strategy is that the index fund have to decrees when the index drops down. In this, situation investors only face with systematic risk [3].

In recent years the passive investment become much more popular based on the following reasons:

- While the best investment funds that follow the active strategy could gain high return than the market return, but most of them failed to beat the market return. Statistical data revealed that 80% of funds which use active strategy couldn't satisfy even their benchmarks [4].
- There is no guarantee for an active fund – a fund that invest actively- to outperform the market in the next year if it has gained more return than market return in the previous year [5].
- As stock markets value gradually grow in long term by passive investment the fund avoid accepting extra risk and gain the market return [3]

In this paper, we are going to make an index fund - a portfolio that reproduces the performance of a stock market index-.An approach to make such a portfolio is to invest on all index securities which called full replication. While it is the easiest way to make index fund but have some problems:

Corresponding author: Bahareh Dadgar, Master of Science in Financial engineering, Department of Industrial Engineering and Management Systems, University of Amirkabir (Tehran Polytechnique), Iran. Baharedadgar@aut.ac.ir

- Some shares that are in the index have very small proportions. Usually it is happen in those indexes which consist of many shares.
- The consisting stocks of the index may change over time so the index fund have to be revised as the index changes.
- It forces high transaction cost to the index fund [3].

In order to avoid above problems, we are going to divide the index stocks into restricted clusters and invest into representation of each cluster instead of full replication. There are three different approaches in modelling the index tracking problem as follows:

- Models based on the Markowitz model: In these models, the aim is to minimize variance of the difference between index return and tracking portfolio return. Hodges was the first one who used the Markowitz model and compared index tradeoff curve with tracking portfolio trade off curve [6].
- Factor based models: in these models, each stock is related to one or more economic factors and the model shows the relationship between them. Chen Jie et al. [7] minimize the variance of the portfolio with constraints on factor betas and expects excess return in their work. Rudd proposed a single factor model to track S&P500 with heuristic [7]. After that, some other researchers like Corielli expanded the model to multiple factor model [8].
- Independent models: Roll[9] used a quadratic function to minimize the tracking error [9]. Frino proposed a model and considered transaction cost and stock dividend too [10]. Beasly used an evolutionary algorithm for index tracking problem [3].

Independent models are still in center of attentions, Okaya et al. [11] used constraint aggregation (CA) technique for the first time in index tracking problem and compared it with the normal solution. However, both solutions led to a same result but using CA reduced the time cost much more [11]. Canakgoz have presented a more advanced model based on the regression. He used a mixed integer linear programming for index tracking and enhanced index tracking which means obtaining more profit than market return [12].Barroet al. [13] presented a multistage tracking error model and solve it in stochastic programming framework. They also considered about transaction cost and liquidity components in their model [13]. Christian et al described a stochastic optimization technique based on the time series clustering for index tracking problem. First, they selected a subset of stocks from the market to construct the tracking portfolio and in the second stage, they used stochastic optimization to weight the selected stocks. Finally, they showed that using clustering enhanced the results [14].

Today's financial analysis depends on mathematical tools and techniques like signal processing which has mathematical base. Andrew Back used blind source separation to extract stocks trend for the first time. He considered the price as a signal and studied the market by independent component analysis (ICA) [15]. Drakakis supposed that each stock price consist of a set of latent variables. He have clustered stocks by non-negative matrix factorization that is one of the signal processing methods [16].

The rest of this paper is organized as follows: In Section 2, the main model is presented. In section 3, ICA and similarity measures are highlighted. Section 4, is dedicated to data. Finally the results are shown in section 5.

2. PORTFOLIO SELECTION MODEL

As presented by Cornuejols and Tutuncu (2007), the model selects stocks for a tracking portfolio. After solving the model, the selections are weighted based on their market value. There are numerous measures of similarity between assets. Cornuejols and Tutuncu used correlation as the similarity measure. Based on their work, we have proposed a model which uses information coupling based on ICA as the similarity measure and compared the results of the two models.

$$\rho_{ij} = \text{similarity measure between stock i and stock j} \tag{1}$$

Suppose we construct a portfolio of q assets from a target index of n assets. Let ρ_{ij}be the similarity between asset i and asset j.Lety_j represent if asset j is selected to be in the portfolio (1 if true, 0 otherwise). Let x_{ij}represent whether asset j is a representative of stock i; x_{ij} is 1 if j is the most similar asset in the portfolio to i, 0 otherwise.

$$Z = \max \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} x_{ij}$$

Subject to : $\sum_{j=1}^{n} y_j = q$ (portfolio size constraint)

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad \text{for } 1 = 1, \dots, n$$

(each stock has exactly one representative in the portfolio)

$$x_{ij} \leq y_j \qquad \text{for } 1 = 1, \dots, n; j = 1, \dots, n$$

(stock must be in the portfolio to be a representative)

$$x_{ij}, y_j = 0 \text{ or } 1 \qquad \text{for } 1 = 1, \dots, n; j = 1, \dots, n$$

After solving this model, a weight ω_j is calculated for each selected asset j using the sum of the market value, V_i of each stock from the index it is representing.

$$\omega_j = \sum_{i=1}^{n} V_i x_{ij} \tag{2}$$

$$\frac{\omega_j}{\sum_{f=1}^{n} \omega_f} \tag{3}$$

3. INDEPENDENT COMPONENT ANALYSIS

In recent years, by exponential increase in the availability and use of financial market data, extracting useful information about similarity between time series become crucial. These large and noisy data set are highly non-Gaussian in nature and require the use of efficient and accurate interaction measurement approaches for their analysis in a real-time environment. There are several approaches to measure similarity but they have some limitations [19]. We summarize these limitations in Table 1.

Table 1. Summary of limitations of similarity measures

Approaches	Limitations
Linear correlation	- Accurately measure only for Gaussian distributions. - Very sensitive to outliers
Rank correlation	- Only valid for monotonic functions - Not suitable for analyzing financial return data - loss of information from the data because of using rank
Mutual information	- Sensitive to multivariate joint distributions estimation and difficult to estimate these distribution in large data base.
Copula	- Complex computational - Sensitive to the empirical choice of the type of copula.

In this paper, we present a new similarity measure based on independent component analysis (ICA) in multivariate non-Gaussian data streams.

ICA is a blind source separation tool which makes use of higher-order statistics and is therefore able to capture information in the tails of multivariate distributions. According to its classical definition, ICA estimates an un-mixing matrix such that the mutual information between the independent source signals is minimized [19]. Hence, we can consider the un-mixing matrix to contain information about the degree of mutual information between the observed signals.

Consider a set of N observed signals $X = [x(t)]_{t=1}^{t=T}$ at the time instant t, which are a mixture of M source *signalss* $= [s(t)]_{t=1}^{t=T}$, mixed linearly using a mixing matrix A, with observation noise n(t), is given by:

$$X(t) = As(t) + n(t) \tag{4}$$

Independent component analysis (ICA) attempts to find an un-mixing matrix W, such that the M recovered source signals $b = [b(t)]_{t=1}^{t=T}$ are given by:

$$b(t) = W(X(t) - n(t)) \tag{5}$$

For the case where observation noise n(t) is assumed to be normally distributed with a mean of zero, the least squares expected value of the recovered source signals is given by:

$$\hat{b}(t) = WX(t) \tag{6}$$

Where W is the inverse of A; that is,

$$W = A^{-1} \tag{7}$$

We can consider the un-mixing matrix to contain information about the degree of mutual information between the observed signals. Equation (8) shows how to calculate information coupling, as a proxymeasure for mutual information [19].

$$\acute{\eta} = \frac{||W||_2 - 1}{\sqrt{N} - 1} \tag{8}$$

Where W is un-mixing matrix that output of Icadec[1] algorithm, $||\cdot||_2$ is the spectral norm of the matrix and N is the number of observed signals. We use information coupling to improve the model as we described above.

4. RESULTS

In order to implement the model we used NASDAQ market data set to track NASDAQ100 index which consists of 100 different stocks. We used log-return of these stocks for 1008 days. 708 records are used to run the model and the other 300 records are used for testing the model. We omit 6 stocks out of 100 stocks due to lack of records (less than

[1]Icadec is one of the ICA algorithm that describe in Appendix 1

1008days). For using Information Coupling criteria the data don't have to be Gaussian distribution so we used the Jarque-bera test in 5% confidence level to find stocks with Gaussian distribution. One other stock is omitted due to its Gaussian distribution. So, the model has been run with 93 stocks data.

5. CONCLUSION

We run the model with both similarity measures correlation and information coupling in order to build the tracking portfolio by investing on just 5 stocks out of 93 available stocks. We choose variance of difference between index return and tracking portfolio return as a measure of error like Beasly (2003) [3]. In fact if the difference between index return and tracking portfolio return remain constant, the error would be zero. In figure 1, index return is shown by red * and the tracking portfolio is shown with a blue line. Figure 2 shows the excess return.

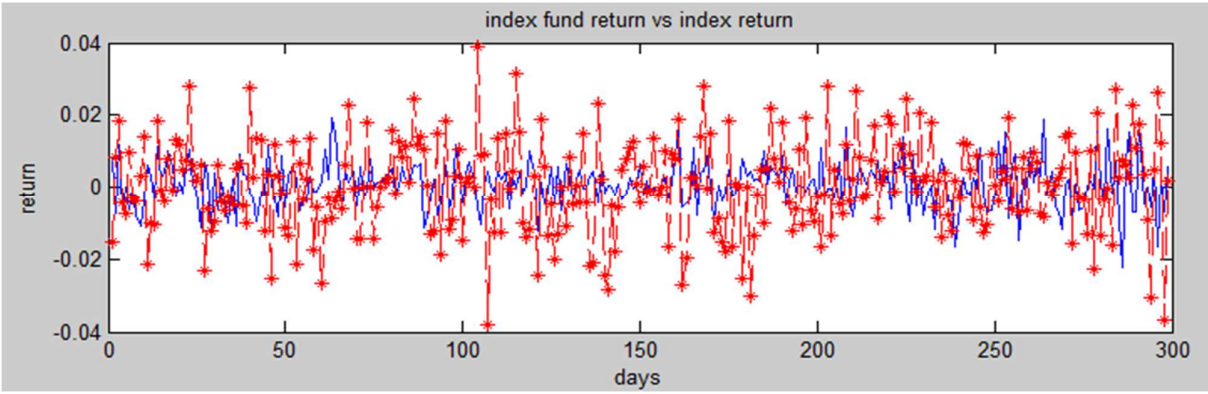

Figure 1. Index return versus tracking portfolio return.

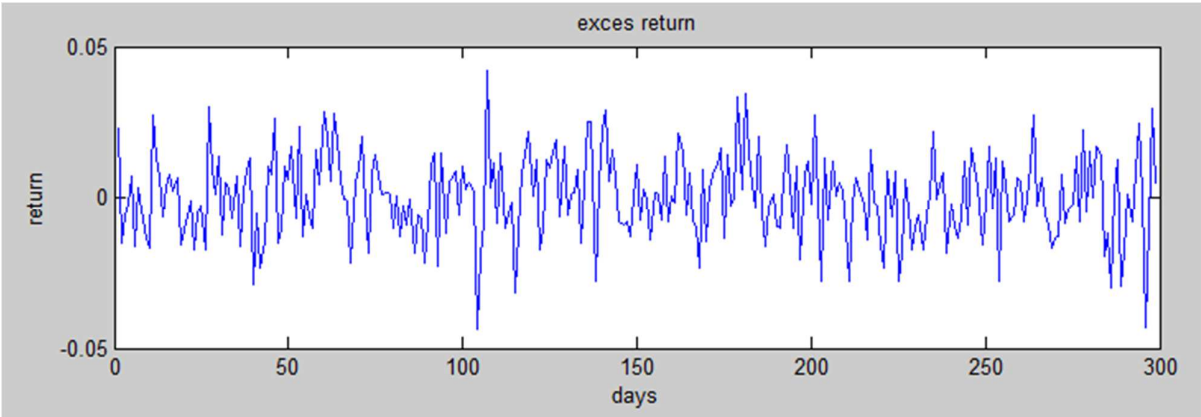

Figure 2. Excess return.

As it is shown in table 1 tracking portfolio error is less when we use information coupling measure in comparison to use correlation measure. As stock returns follow a non-Gaussian distribution using information coupling can be very useful in this field.

Table2. Results of the model based on two different similarity measure.

Similarity measure	Correlation	Information coupling
Error	0.0097	**0.0084**

REFERENCES

1. b. P. H. Hogan, "Portfolio theory creates new investment opportunities," Journal of Financial Planing, pp. 35-37, January 1994.

2. W. F. Sharpe, "Capital asset prices: a theory of market equilibrium under conditions of risk," The Journal of Finance, vol. 19, no. 3, p. 425–442, September 1964.

3. J.E. Beasley, N. Meade, T.-J. Chang, "An evolutionary heuristic for the index tracking problem," European Journal of Operational Research, p. 621–643, 2003.

4. Y. W. He Ni, "Stock Index tracking by Pareto efficient genetic algorithm," Journal of Applied Soft Computing 13, p. 4519–4535, 2013.

5. J. C. Bogle, "Selecting equity mutual funds," The Journal of Portfolio Management , pp. 94-100, 1992.

6. S. D. Hodges., "Problems in the application of portfolio selection models.," Omega, Volume 4, Issue 6, p. 699–709, 1976.

7. Andrew Rudd, "Optimal Selection of Passive Portfolios,," Financial Management, Vol. 9, No. 1 ,, pp. 57-66, 1980.

8. M. M. Francesco Coriellia, "Factor based index tracking," Journal of Banking & Finance, vol. Volume 30, no. Issue 8, p. 2215–2233, 2006.

9. R. Roll, "A mean/variance analysis of tracking error.," The Journal of Portfolio Management 18 (Summer), , pp. 13 - 22 ., 1992 .

10. D. R. ,. G. Alex Frino, "Tracking S&P 500 Index Funds,," The Journal of Portfolio Management , Vol. 28, No. 1, pp. 44-55, 2001.

11. ,. Nesrn Okaya & Uğur Akmanb, "Index tracking with constraint aggregation,," Applied Economics Letters , pp. Volume 10, Issue 14, 2003.

12. N.A. Canakgoz, "Mixed-integer programming approaches for index tracking and enhanced indexation," European Journal of Operational Research 196 (2118), p. 384–399, 2009.

13. E. C. Diana Barro, "Tracking error: a multistage portfolio model," Annals of Operations Research, pp. Volume 165, Issue 1, pp 47-66, 2009.

14. S. C. Christian Dose, "Clustering of financial time series with application to index and enhanced index tracking portfolio," Physica A , p. 145–151, 2005.

15. A. S. W. Andrew D. Back, "A First Application of Independent Component Analysis to Extracting Structure from Stock Returns.," NYU , p. working paper, 1997 .

16. S. R. R. d. F. A. C. Konstantinos Drakakis, "Analysis of Financial Data Using Non-Negative Matrix Factorization," InteInternational Mathematical Forumrnational Mathematical Forum, p. 1853 – 1870, 2008, .

17. T. R. Cornuejols G, Optimization methods in finance, Cambridge University Press., 2007.

18. R. H. K. Chen Chen, "Robust portfolio selection for index tracking,," Computers & Operations Research, p. 829–837., 2012.

19. N. S. &. S. J.Roberts, "Dynamically Measuring Statistical Dependencies in Multivariate Financial Tim Series Using Independent Component Analysis," ISRN Signal Processing, vol. Volume 2013, 2013.

20. R. E. a. S. Roberts, "Independent component analysis: a flexible nonlinearity and decorrelating manifold approach," Neural Computation, vol. 11, p. 1957–1983, 1999.

Appendix.

In this section, we explain overview of Icadec algorithm. More details presented in [19].

For a set of observed signals X, where $X = [x(t)]_{t=1}^{t=T}$, the set of recovered independent components,$B = [b(t)]_{t=1}^{t=T}$,is given by B = WX. The independent components are linearly de-correlated if Equation (9) satisfy.

$$\mathrm{BB^T} = \mathrm{WXX^TW^T} = \mathrm{D^2} \tag{9}$$

Where D is a diagonal matrix of scaling factors. The singular value decomposition of the set of observed signals is given by:

$$\mathrm{X} = \mathrm{U\Sigma V^T} \tag{10}$$

It can be shown that the un-matrix, W can then be written as:

$$\mathrm{W} = \mathrm{DQ\Sigma^{-1}U^T} \tag{11}$$

Where Q is a real orthogonal matrix. To obtain an estimate for the ICA un-mixing matrix, we need to optimize a given contrast function. We use log-likelihood of the data, as described below.

Assuming that the observation noise is normally distributed with a mean of zero and having an isotropic covariance matrix with precision β, the distribution of the observations conditioned on A and s is given by:

$$\mathrm{p(x|A,s)} = \mathrm{N\left(x; As, \beta^{-1}I\right)}, \tag{12}$$

Where As is the mean of the normal distribution and $\beta^{-1}I$ is its covariance. The likelihood of an observation occurring is given by:

$$\mathrm{p(x|A)} = \int \mathrm{p(x|A,s)p(s)ds}. \tag{13}$$

Assuming that the distribution over sources has a single dominant peak, in this case given by the maximum likelihood source estimates $\hat{s} = (A^TA)^{-1}A^TX$, the integral in (11) can be analyzed by using a simplified (computationally efficient) variant of Laplace's method, as shown in Equation (14)

$$\mathrm{p(x|A)} = \int \mathrm{p(x|A,s)p(s)ds} \approx \mathrm{p(x|A,\hat{s})p(\hat{s})(2\pi)^{M/2}\,det(G)^{-1/2}} \tag{14}$$

Where G is a hessian Matrix:

$$\mathrm{G} = -\left[\frac{\partial^2 \mathrm{log p(x|A,s)}}{\partial \mathrm{s_i}\,\partial \mathrm{s_j}}\right]_{\mathrm{s=\hat{S}}} = \mathrm{\beta A^TA} \tag{15}$$

A is inverse of W so we have: A=U$\Sigma Q^T D^{-1}$

The log-likelihood, is Therefor:

$$\log \mathrm{p(x|A)} = \frac{\mathrm{N-M}}{2}\log\left(\frac{\beta}{2\pi \mathrm{e}}\right) + \log \mathrm{p(\hat{s})} + \log\det\left(\Sigma^{-1}\mathrm{D}\right) \tag{16}$$

Following the resulting likelihood gradient using a Broyden-Fletcher-Golfarb-Shanno (BFGS) optimizer, makes it possible to efficiently compute an optimum value for the ICA un-mixing matrix.