Searching Co-Integrated Portfolios by a Genetic Algorithm

Pravesh Kriplani
Brics Securities Lmt.
Quantitative Strategy Group
Sadhana House, 1st Flr, 570
400018 Mumbai – India
pravesh.kriplani@bricssecurities.com

Luigi Troiano
University of Sannio
Department of Engineering
RCOST – Viale Traiano
82100 Benevento – Italy
troiano@unisannio.it

***Abstract*—Searching for portfolios co-integrated with an index offers new opportunities in designing robust investment strategies. The problem of finding optimal index co-integrated portfolios that are maximally stationary is combinatorial. Indeed, given a basket of equities, the portfolio/index co-integration cannot be simply expressed in terms of equity/index co-integration. In this paper we investigate the application of simple genetic algorithms in finding optimal portfolios.**

I. Introduction

Changing dynamics of the financial markets are driving factors behind the search for robust portfolio strategy. One of the key factors for such strategies is the ability of the coupled and self-consistent optimization of the strategies. Another is to be interpretable and provide robust built-in complexity controls that ensure acceptable out-of-sample performance [1]. However, the increasing number of assets in the financial markets, and increasing complexity of investment environments depending on business cycles and more integrated global markets leads to time-invariability in the performance of funds [2].

This more than ever let us to build a portfolio before actionable information deteriorates. Novel approaches have been done in the discovery of trading strategies for co-integrated instruments, but in an ever increasing space of assets the search is getting more computationally intensive, resulting in the use of qualitative stock selection procedures [3].

Existing approaches of quantitative portfolio strategies rely on correlation assumptions. This approach has a number of shortcomings, amongst which instability is most hazardous. Correlation is valid only for stationary variables. This requires prior de-trending of prices, however, this procedure has the disadvantage of loosing valuable information. By de-trending the variables before analysis removes any possibility to detect common trends in prices.

By contrast, according to Alexander [2] the aim of the co-integration analysis is to detect any stochastic trend in the price data and use these common trends for a dynamic analysis of correlation in returns. Also when portfolios are constructed based on correlation, it is necessary to rebalance them frequently, while co-integrated portfolios might deviate in the short run but they should be tied in the long run. Optimal co-integration portfolios should therefore not require much rebalancing.

Discovering optimal co-integration portfolios entails to consider different combinations of assets in order to find that best follows a market benchmark. Because of joint statistical effects in price series, co-integration of the whole (portfolio) cannot be obtained as co-integration of the parts (assets), entailing combinatorial complexity in solving the problem. This suggests that genetic algorithm can play a role in approaching a solution. The reminder of this paper is organized as follows: Section 2 provides some preliminaries on co-integration analysis; Section 3 outlines a solution by genetic algorithms; Section 4 presents some preliminary experimental results; Section 5 is devoted to conclusions and future directions.

II. Related work

Application of genetic algorithms to portfolio optimization is not new to scientific and financial community. In literature several experiences are reported. They generally refer to optimizing portfolios in terms or risk adjusted returns.

Chi-Cheong [4] makes use of Genetic Algorithms to optimize asset allocation in terms of profit maximization at predefined risk level. In order to avoid premature convergence, new schema are injected into population with random generation of new strings when convergence occurs, with the double aim of improving space exploration by keeping high population genetic variety and to save computational time due to similar but not excellent solutions. Injection of new schema is performed by two procedures: full reshuffling and partial replacement. The first, restarts the population from the beginning running actually a new algorithm in sequence. The second takes advantage of the fact that some good attributes might have been acquired previously by some strings through the reproduction process. Solution is tested by optimizing cash and various stocks in Hong Kong market over extended time periods.

However risk/return optimization is inherently multi-objective. Coello Coello [5] provides an overview of literature. Generally solutions share the approach in identifying the Pareto's frontier offering a set of solutions from high-return/high-risk to low-return/low-risk.

978-1-4244-5612-3/09/$26.00 ©2009 IEEE

Hassan et al. [6] argue that identifying the return/risk frontier is not enough as solutions could evolve as the environment changes, and the relative positions of previously identified solutions may alter. So low-risk solutions may become high-risk and vice versa. They investigate the problem by proposing a Multi-Objective Genetic Programming (MOGP) algorithm such as Strength Pareto Evolutionary Algorithm 2nd version (SPEA2) and including a new robustness measure into a MOGP fitness function to bias evolution towards more robust solutions.

Portfolio optimization include additional constraints, such as cardinality constraints, buy-in thresholds, roundlots etc. Streichert et al. [7] consider genetic algorithms jointly to local search for exploring feasible solutions. Skolpadungket et al. [8] experimented various techniques of multi-objective genetic algorithms to solve portfolio optimization with some realistic constraints, namely cardinality, floor and round-lot. They consider Vector Evaluated Genetic Algorithm (VEGA), Fuzzy VEGA, Multi-objective Optimization Genetic Algorithm (MOGA), SPEA2 and Non-Dominated Sorting Genetic Algorithm 2nd version (NSGA2), comparing their performances. In this paper we only consider the cardinality constraint, looking for solutions with a minimum and maximum number of equities to be considered. We share with others, the coding of solutions as an array of integers, each indexing an equity. Therefore the maximum cardinality is imposed by the chromosome length. In addition co-integrating vectors tend to use the maximum cardinality, as when this is much below the basket size (e.g. 10 over 100 and more) it is very likely to find a co-integrated solution assuming the maximum number of equities.

Aranha and Iba [9] offer an overview of other codings, entails arrays of reals for portfolio weights, or binary, where each element represents the presence or absence of an asset in the portfolio. In addition they propose a tree structure to represent a portfolio where intermediate nodes represent the weights, and the leaves represent the assets.

III. Portfolio Co-Integration

The term *co-integration*, a formulation of the phenomenon that non-stationary processes can have linear combinations that are stationary, was coined by Granger [10].

Stock and Index levels are usually non-stationary, the level of a stock or indices can be postulated as an accumulated process of some stationary process. In fact, if we resort to the statistical hypothesis testing procedures (unit root tests[1]), most of the stock and index levels are judged to be non-stationary. Thus, it is not easy to build a model with satisfactory forecast accuracy because the forecast error variance diverges rapidly with the increase of forecast horizon [11]. The story changes if we have a long short portfolio where subtracting the price of Portfolio A from that of Portfolio B can lead to seemingly a stationary series around a fixed mean.

[1] *Unit root tests* is a class of statistical procedures aimed at testing non-stationary of time series by auto-regressive (AR) models.

Stationarity. A time series is said to be *stationary* if it has a constant mean, variance and a covariance that only depends on the time between lagged observations. Stationary time series are integrated of order zero or $I(0)$. Order of integration $I(p)$ is the minimum number of differences required to obtain a stationary series. An important property is that a $I(p)$ series can be constructed by summing (integrating) a $I(p-1)$ series, or conversely differentiating a $I(p)$ series leads to a $I(p-1)$ series.

Unfortunately most stock prices are not stationary as they exhibit a geometric random walk that gets them farther away from their starting values. We would not expect if any time series of prices, exchange rates or index levels to be stationary because of rising and falling levels are a feature of financial variables. However, the returns required by investors should be dependent upon the uncertainty surrounding the investment and independent of the levels of index or stocks. Thus returns data may have a constant mean and standard deviation and a covariance between observations that depends only upon the number of lags between those observations [12]. In other terms we should assume returns of order $I(0)$, thus prices being $I(1)$.

Co-Integration. Referring to the notion of co-integration as first formulated by Engle and Granger[10], two time series $X_t \sim I(1)$ and $Y_t \sim I(1)$ are co-integrated if there exists a linear combination $Y_t - \beta X_t = Z_t$ such that $Z_t \sim I(0)$. β is referred to as the coefficient of co-integration.

Regression models between co-integrated series are not spurious, therefore if a relationship between stochastic variables exists this can be assumed with high confidentiality.

Co-integration plays a relevant role in linking futures and spot in statistical arbitrage. If futures is above its fair value arbitrageurs would sell the future and buy the index, forcing the prices back to their equilibrium level. Conversely, if the future were below its fair value, arbitrageurs would buy the futures and sell the index. Lets assume that a fair futures price of $\beta = 1.1$ times the equity index level is equilibrium relationship. This can be expressed as $Y_t = 1.1X_t$. This is the same as saying $Y_t - 1.1X_t = 0$. This relationship only holds in equilibrium. In the short run each of these variables will vary over time in its own way. If there is a long-run equilibrium relationship between them, $Y_t - 1.1X_t = Z_t$. Thus if an equilibrium relationship exists Z is stationary.

There exists at least three methods for testing co-integration: Eagle-Gender [10], Phillips-Ouliaris [13] and Joahnsen [14].

Portfolio Selection and Strategies. As opposed to long-short strategies, market neutral strategies involve only equities or securities with proved interdependencies. Such interdependencies, sometimes take the form of convergence, ensuring that, over a given time horizon, the equities will reach an assumed pricing relationship.

Co-integration can be extended to multiple variables and used to construct portfolios. Portfolios constructed using co-integration is an example of multivariate co-integration. For

example let us consider series X_t, Y_t, W_t. We can find a regression model such that $aX_t+bY_t+cW_t = Z_t$ is stationary. Among variables we will include a stock market index in order to link the portfolio to it.

Index Tracking is a portfolio strategy where the aim is to replicate the benchmark in terms of returns and volatility using co-integration. This allows us to make use of the full information contained in stock prices and base portfolio weights on the long run behavior of stocks. This implies creating a basket of securities which replicate the index while taking care of the transaction and impact cost. To ideally form a portfolio to replicate the index should not exceed 20%.

Long-Short market neutral strategy consist of buying a portfolio of attractive stocks, the long portion of the stock, and selling a portfolio of unattractive stocks, the short portion of the portfolio. The spread between the performance of the two provide a return in excess of the risk free rate, because of the significantly higher risk and return expectation.

In practice the construction of both *long* and *short* portfolios can be derived from the index tracking strategy; only this time we aim to devise two co-integrating portfolios tracking 2 benchmarks, a benchmark B^+ and a benchmark B^- constructed by adding to (respectively subtracting from) the main benchmark daily returns an annual excess return expressed in percentage (equally distributed on the daily returns) as suggested by Dunis [3].

Finally as the long and short portfolios are both highly correlated with the reference stock market benchmark and assuming that each tracking error is not correlated with the market, one would expect a low correlation of their difference with the market benchmark, a key characteristics of a market neutral strategy.

IV. Searching optimal portfolios

Since the number of constituents in a diversified 10-equities portfolio like NIFTY are 50, the possible combinations are $\binom{50}{10}$ to consider all future stocks would be $\binom{183}{10}$. Therefore the optimization of asset allocation in a portfolio is complex, and NP-hard. Meta-heuristics such as genetic algorithms can provide a feasible way to find near-optimal solutions.

Algorithm. In this paper, we preliminarily considered a Simple GA, as proposed by Goldberg [15], entailing a sequence of three genetic operators: selection, crossover and mutation. An initial portfolio population P_0 is created by uniformly choosing equities in the basket and evaluated. After, evolution goes ahead until $genlimit$ is reached. At each generation i, a new population P_i is produced by selecting and reproducing individuals from previous population P_{i-1}. For selecting individuals we chose 2-*tournament* operator as more robust to fitness scaling. Pairs of contiguous parents $< h, h+1 >$ are crossed, and individuals h and $h+1$ mutated at given crossover and mutation rate. For crossing individuals we adopted *scattered crossover*: a number of genes is randomly chosen and exchanged between parents. For mutation, we simply choose a gene and we alter its value. If average fitness value does not change significantly for a given number of generations, fitness is stalled and algorithm quited.

Algorithm 1 Genetic Algorithm.

```
1: P_0 ← Random(popsize)
2: Evaluate(P_0)
3: for i = 1..genlimit do
4:     P_i ← Select(P_{i-1})
5:     for all < h, h+1 > ∈ P_i do
6:         Cross(< h, h+1 >), if selected for crossover
7:         Mutate(h), if selected for mutation
8:         Mutate(h+1), if selected for mutation
9:     end for
10:    Evaluate(P_i)
11:    if fitness stalled then
12:        Exit
13:    end if
14: end for
```

Solution coding. A portfolio is represented by indexes reporting equities in the basket, as depicted in Fig.1. Therefore genes are integers between 0 (meaning a null pointer) and the basket size (i.e. the index of last equity in the basket). Equities can be reported more than once, thus avoiding to consider the no-repetition constraint. Portfolio maximum cardinality is implicitly imposed by the chromosome length.

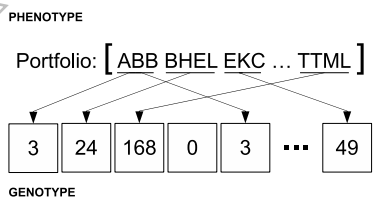

Fig. 1. An example of solution

Fitness function. Each solution is evaluated according to the following 3-steps procedure:

1) Test portfolio/index co-integration by Joahnsen test.
2) If they result co-integrated, retrieve the most likely co-integrating vector
3) and test AR model stationarity by augmented Dickey-Fuller test (ADF).

The use of OLS regression is not suitable for identifying the various co-integrating vectors which may exist within the multivariate framework. A more powerful test of the components of the co-integrating vector is the Johansen likelihood ratio or *trace* test, which draws on the error correction model to specify the independent co-integrating vectors and to test for their stationarity. In our experimentation we considered the trace statistics with confidence level of 99%.

Algorithm goal is to find a co-integrated portfolio which is maximally stationary. The ADF statistic is a negative number: the more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence. The genetic algorithm is aimed at discovering those solutions that minimize the ADF statistic. Ultimately the fitness value provided is the ADF statistic, if Joahnsen and ADF null hypotheses are rejected, 0 otherwise.

V. Experimental results

In order of experiment our approach we considered a basket of futures listed at the Indian National Stock Exchange (NSE). The Indian Equities and Derivatives Markets is one of the largest growing in terms of volume and value trades. India has the distinction of having the second largest number of listed companies after the USA. According to Standard and Poors Fact Book 2008, NSE is the most active exchange in India in terms of turnover. NSE is the third largest Stock Exchange in the world in terms of the number of trades in equities. In 2008, the equity market capitalization of the companies listed on the NSE was USD 1.819 trillion, making it among the top 10 stock exchange markets in the World for capitalization and trades. Volumes of stock futures more than doubled in recent years.

We considered the 183 equity stocks traded at futures exchange in the period 6th Aug 2008 - 4th Aug 2009. So each time series was made of 240 points representing the daily closing price expressed in Indian rupees (INR). As benchmark, we considered the NIFTY index. The S&P CNX NIFTY, a free float market capitalization index, is the leading index for large companies on the National Stock Exchange of India. It consists of 50 companies representing 24 sectors of the economy. The base level is defined as 1000 on November 3, 1995. In January 2005, its level was almost 2000.

Solution landscape. The first test was aimed at verifying the genetic algorithm ability to explore the search space in discovering optimal portfolios. Fig.2 depicts the histogram of co-integrated portfolios made of 5 equities, obtained by randomly sampling (354230 points) all possible combinations. Feasible portfolios were 21769. In Fig.3 we report the fitness

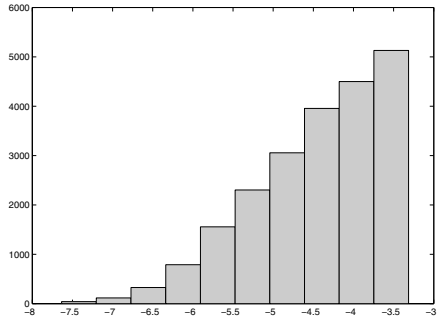

Fig. 2. Distribution of 5-equities portfolio fitness

histogram of best portfolios discovered in 100 simulations of the genetic algorithm with standard crossover rate 0.8, generation limit 100, population size 100, mutation probability 0.02. We notice genetic algorithm was able to get a solution among the top 5% portolios (i.e. 1088) in 89% of simulations, and among the top 1% (i.e. 218) in 54% of cases. The worst solution provided by the genetic algorithm was in the first quartile of feasible portfolios, median in top 1%.

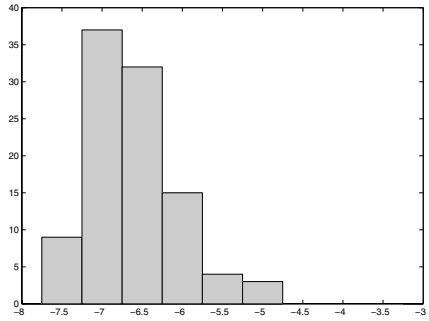

Fig. 3. Distribution of ga best individuals fitness

Genetic Algorithm vs. Random Search. In this experiment we compared the performance of genetic algorithm versus random search in finding optimal portfolios made of 5 equities. In order to make comparison fair, random search was iterated over 100 generations, each considering 100 random solutions. At each iteration, the best individual is kept. Similarly, genetic algorithm was run with generation limit 100, population size 100. Other parameters were left unchanged to the previous case. Fig.4 outlines the average fitness of best solutions obtained along 20 runs of the experiment.

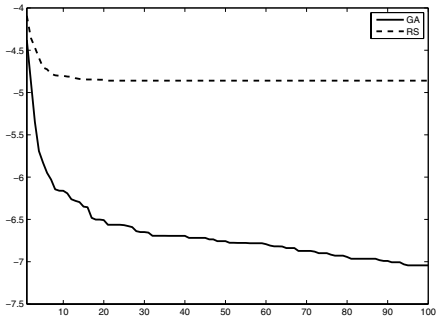

Fig. 4. Fitness evolution of GA vs. RS

Larger portfolios. We considered how the algorithm performs when larger portfolios are considered. In Fig.5 we report the average of best solution fitness (along 20 runs) at different portfolios size (i.e. 10, 15, 20). In this case we set the generation limit at 200. We also performed a Kruskall-Wallis test and unpaired Wilcoxon test in order to establish if there was a statistical difference at the end of generations. In both cases the null hypothesis of belonging to the same population was rejected at confidence level of 95%. This is in accordance with the common understanding that larger prortfolios are more stably related to the stock marked index. Fig.6 provides the histogram of best solutions as provided by the algorithm along the experiment.

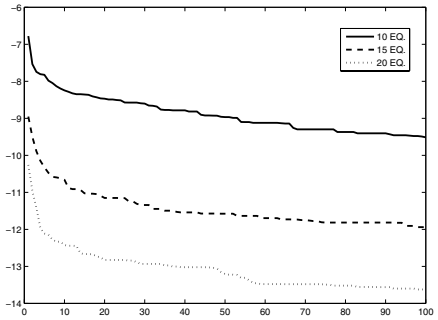

Fig. 5. Portfolio size. Average fitness of best individuals.

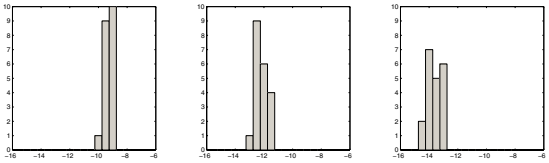

Fig. 6. Distribution of ga best individuals fitness

Population Size. The number of individuals considered at each generation can affect the algorithm convergence. In this experiment we compared the algorithm with populations made of 50, 100 and 200 individuals. The average evolution of best individuals is plotted in Fig.7. Instead Fig.8 provides a scatter plot of average fitness evolution. As expected, we note that evolution performs better in the case of larger populations.

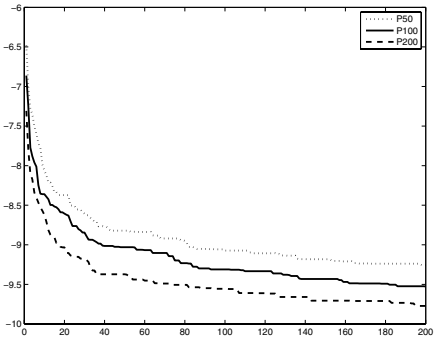

Fig. 7. Population size. Distribution of ga best individuals fitness.

VI. Conclusions and future work

In this paper we have investigated the application of genetic algorithms in searching optimal co-integration portfolios. The work is still at beginning and we can only draw some preliminary conclusions. The main outcome is that genetic algorithms, even in their simpler form, can effectively face the problem of discovering near-optimal portfolios co-integrated with a stock market index. Many questions arise regarding performances and convergence. For instance how far it is possible to enlarge the portfolio in order to consider a wider set of equities. Moreover, we often observed a genetic drift resulting in an early convergence of the algorithm. This suggest to implement strategies in order to make the algorithm restarting. Another interesting direction regards the application of multi-niche oriented genetic algorithms, able to consider a set of sub-optimal portfolios at once.

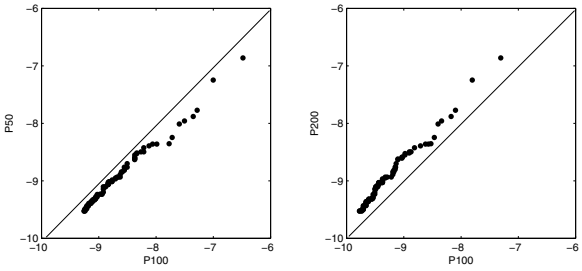

Fig. 8. Population size. Scatter plot.

References

[1] V. V. Gavrishchaka, "Discovery of multi-spread portfolio strategies for weakly-cointegrated instruments using boosting-based optimization," in *JCIS*. Atlantis Press, 2006.

[2] C. Alexander and A. N. C. A. Dimitriu, "The cointegration alpha: Enhanced index tracking and long-short equity market neutral strategies," *Social Science Research Network Working Paper Series*, June 2002.

[3] C. L. Dunis, "Cointegration portfolios of european equities for index tracking and market neutral strategies," *Journal of Asset Management*, vol. 6, no. 1, pp. 33–52, June 2005.

[4] Chi-Cheong, "Genetic algorithms in portfolio optimization," Society for Computational Economics, Computing in Economics and Finance 2001 204, Apr. 2001.

[5] M. G. C. Tapia and C. A. C. Coello, "Applications of multi-objective evolutionary algorithms in economics and finance: A survey," in *IEEE Congress on Evolutionary Computation*. IEEE, 2007, pp. 532–539.

[6] G. Hassan and C. D. Clack, "Multiobjective robustness for portfolio optimization in volatile environments," in *GECCO '08: Proceedings of the 10th annual conference on Genetic and evolutionary computation*. New York, NY, USA: ACM, 2008, pp. 1507–1514.

[7] F. Streichert, H. Ulmer, and A. Zell, "Evolutionary algorithms and the cardinality constrained portfolio optimization problem," in *Operations Research Proceedings 2003, Selected Papers of the International Conference on Operations Research (OR 2003*. Springer, 2003, pp. 3–5.

[8] P. Skolpadungket, K. Dahal, and N. Harnpornchai, "Portfolio optimization using multi-objective genetic algorithms," in *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*, Sept. 2007, pp. 516–523.

[9] C. C. Aranha and H. Iba, "A tree-based ga representation for the portfolio optimization problem," in *GECCO '08: Proceedings of the 10th annual conference on Genetic and evolutionary computation*. New York, NY, USA: ACM, 2008, pp. 873–880.

[10] R. F. Engle and C. W. J. Granger, "Co-integration and error correction: Representation, estimation, and testing," *Econometrica*, vol. 55, no. 2, pp. 251–76, March 1987.

[11] Y. Kawasaki, S. Tachiki, H. Udaka, and T. Hirano, "A characterization of long-short trading strategies based on cointegration," in *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, 2003, pp. 411–416.

[12] T. J. Watsham and K. Parramore, *Quantitative Methods in Finance*. Intl. Thomson Business Press, Sep. 1996.

[13] P. C. B. Phillips and S. Ouliaris, "Asymptotic properties of residual based tests for cointegration," *Econometrica*, vol. 58, pp. 165–193, 1990.

[14] S. Johansen, "Statistical analysis of cointegration vectors," *Journal of Economic Dynamics and Control*, vol. 12, no. 2-3, pp. 231–254, 1988.

[15] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, January 1989.