COMISEF Computational Optimization Methods in Statistics, Econometrics and Finance

- Marie Curie Research and Training Network funded by the EU Commission through MRTN-CT-2006-034270 -

COMISEF WORKING PAPERS SERIES

WPS-036 17/05/2010

Asset Allocation under Hierarchical Clustering

J. Zhang
D. Maringer

Asset Allocation under Hierarchical Clustering

Jin Zhang[*] Dietmar Maringer [†]

Abstract

This paper proposes a clustering asset allocation scheme which provides better risk-adjusted portfolio performance than those obtained from traditional asset allocation approaches such as the equal weight strategy and the Markowitz minimum variance allocation. The clustering criterion used, which involves maximization of the in-sample Sharpe ratio (SR), is different from traditional clustering criteria reported in the literature. Two evolutionary methods, namely Differential Evolution and Genetic Algorithm, are employed to search for such an optimal clustering structure given a cluster number. To explore the clustering impact on the SR, the in-sample and the out-of-sample SR distributions of the portfolios are studied using bootstrapped data as well as simulated paths from the single index market model. It was found that the SR distributions of the portfolios under the clustering asset allocation structure have higher mean values and skewness but approximately the same standard deviation and kurtosis than those in the non-clustered case. Genetic Algorithm is suggested as a more efficient approach than Differential Evolution for the purpose of solving the clustering problem.

Key words. Asset Allocation, Clustering Technique, Sharpe Ratio, Evolutionary Approach, Heuristic Optimization.

1 Introduction

Generally speaking, traditional asset allocation strategies can be classified into two categories: parametric approaches (e.g. the Markowitz allocations) and parameter-free allocations (e.g. the equal weight strategy). According to the well-known Markowitz theory, rational investors should always prefer the 'efficient' portfolios

[*]Centre for Computational Finance and Economic Agents, University of Essex, United Kingdom. jzhangf@essex.ac.uk

[†]Faculty of Economics and Business Administration, University of Basel, Switzerland. dietmar.maringer@unibas.ch

which yield the highest return at any given risk level. However, most of the time, a precise estimation of asset return properties (such as expected return, variance and covariance) may be difficult to obtain. As empirically observed financial data is quite noisy, the estimates derived from such data may be unreliable. Moreover, the accuracy of estimates may rely on not only the number of assets but also the available number of observations, which is particularly important to the Markowitz allocations. If a portfolio contains hundreds of assets, the high dimensionality may hinder an accurate estimation of the dependence structure of assets (i.e. the covariance under the Markowitz framework), which in the literature is usually referred to as the 'curse of dimensionality'. The above problems may result in suboptimal portfolios if investors still apply the traditional asset allocations to manage portfolios, especially large ones. For instance, the Sharpe ratio (SR) of portfolios, which is a risk-adjusted performance measure based on the first two moments of returns, may not be optimal. Many researchers have suggested different approaches for improving the estimation of return moments and portfolio performance. For example, Harris and Yilmaz [2007] combined the return-based and the range-based measures of volatility to improve the estimate of the multivariate conditional variance-covariance matrix. On the other hand, practitioners may simply adopt parameter-free allocations (e.g. the equal weight (EW) investment strategy of Windcliff and Boyle [2004]) which are independent of those return moment measures. In addition to methods from mathematics and finance, approaches from computer science have also been considered by researchers. For instance, Lisi and Corazza [2008] proposed an active fund management strategy which performed portfolio selection after clustering equities. Pattarin et al. [2004] employed a clustering technique to analyze mutual fund investment styles. The clustering techniques considered in most financial applications still comply with the traditional clustering criterion, i.e. minimizing the dissimilarity between the cluster members while maximizing the dissimilarity between clusters.

This paper, however, proposes a different clustering criterion to the traditional one. The proposed clustering criterion segments assets by maximizing the in-sample SR of a portfolio. Two main benefits are expected from using this clustering asset allocation scheme. First, the dimensionality in the asset allocation problem will decrease, as the cluster size can be controlled by using a cardinality constraint. Thus, when portfolio managers apply the Markowitz allocations to managing large portfolios, the 'curse of dimensionality' problem may be avoided. Secondly, the out-of-sample SR of a portfolio which is constructed under the clustering structure shall be better than the portfolio SR from using the same asset allocation in the non-clustered case.

Provided there is no structural break between the in-sample and the out-of-sample periods, and the clustering structure is optimal, one should observe the same clustering impact on both the in-sample and out-of-sample SRs. The first four moments of the SR distribution are considered in this paper in order to study the clustering impact on the SR. A rational investor should prefer the SR distributions with high mean, high skewness, low standard deviation and low kurtosis. To study the four moments of the SR distribution, simulated portfolio returns are generated by using the portfolio weights based on simulated asset returns. The simulated SR values which are calculated from the portfolio returns then constitute the SR distribution. The EW strategy and the Markowitz minimum variance portfolio (MVP) allocation which belong to the non-parametric allocation and the parametric allocation, are adopted to distribute asset weights. Two approaches are used to provide the simulated asset returns for both the in-sample and out-of-sample SR studies, i.e. the traditional bootstrap method and the single index market model.

The proposed clustering asset allocation scheme is introduced starting with the following technical terms. Assets within a cluster are called 'cluster members', the portfolio which is constructed by using the members in the same cluster is referred to as a 'cluster portfolio', and these cluster portfolios are combined to form a 'terminal portfolio'. In other words, the proposed approach first segments assets into a series of disjoint clusters according to the clustering structure. Next, a set of cluster portfolios is constructed by using an asset allocation on the basis of the cluster members in different clusters. Finally, the terminal portfolio is constructed by adopting the same asset allocation based on those cluster portfolios. Figure 1 briefly describes this clustering asset allocation procedure in a case of three clusters with eleven assets.

The paper is organized as follows. Sections 2 and 3 introduce the clustering optimization problem and asset allocation approaches used. Section 4 describes two evolutionary algorithms for tackling the clustering problem. The experimental results are presented and discussed in Section 5. Section 6 summarizes the main findings.

2 The Optimization Problem for Clustering Asset Allocation

Supposed there are N stocks considered in the asset allocation problem. The optimization problem is to identify a clustering structure $\mathcal{C}$ (i.e. a union of subsets $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_G$), so that the portfolio SR based on such a cluster structure is maximized given a cluster number G. In this paper, the cluster number G is manually assigned, and G is an integer number within a range $1 \leq G \leq N$ as empty clusters are not considered. When G is equal to the number of either 1 or N, there is

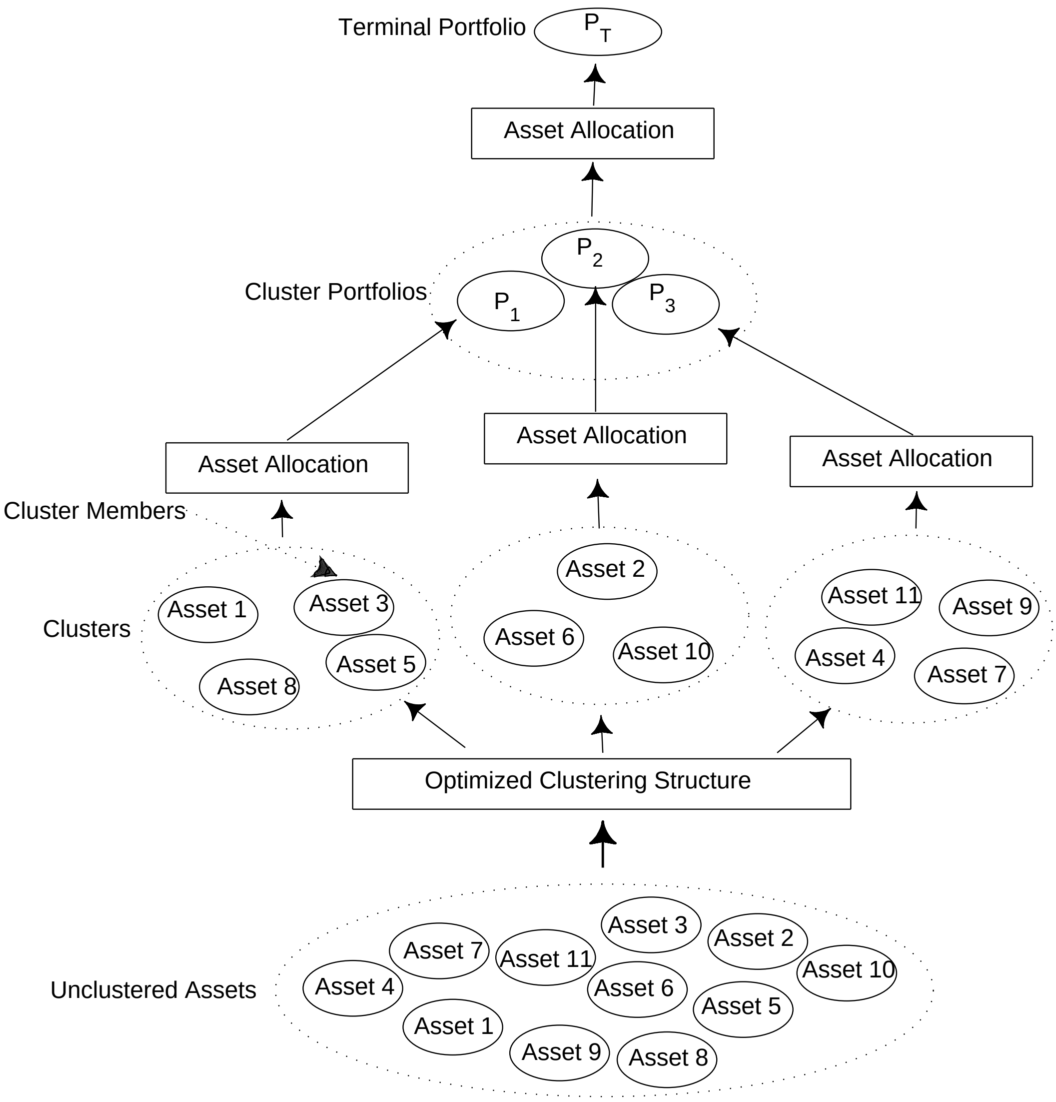

Figure 1: Procedure for the Clustering Asset Allocation

no clustering effect, thus the clustering asset allocation problem becomes a simple application of the traditional asset allocation. The optimization objective of the clustering problem can be described as

$$\max_{\mathcal{C}} SR = \frac{r_P - r_f}{\sigma_P}, \tag{1}$$

where $\mathcal{C}$ denotes the optimal clustering structure, r_P is the average daily return of the portfolio, σ_P is the standard deviation of the portfolio return over the evaluation period, and r_f refers to as the risk-free return. As with traditional clustering problems, the union of segmented assets $\mathcal{U}$ represents the collection of assets, and there is no intersection between two different clusters. Let $\mathcal{C}_g$ denote the g-th cluster of assets, then the above constraints can be written as

$$\bigcup_{g=1}^{G} \mathcal{C}_g = \mathcal{U}, \tag{2}$$

$$\mathcal{C}_g \cap \mathcal{C}_h = \emptyset, \quad g \neq h. \tag{3}$$

Let $\tilde{N}^{\min}$ and $\tilde{N}^{\max}$ denote the minimum and maximum asset numbers allowed in a cluster respectively, then the following cardinality constraints are employed to limit the dimensionality of clusters:

$$\tilde{N}^{\min} \leq \sum_{j=1}^{N} I_{j\in\mathcal{C}_g} \leq \tilde{N}^{\max} \qquad 1 \leq g \leq G, \tag{4}$$

$$\text{where} \quad I_{j\in\mathcal{C}_g} = \begin{cases} 1 & \text{if } j \in \mathcal{C}_g, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

$$\text{with} \quad \begin{cases} \tilde{N}^{\min} = \lceil \frac{N}{2G} \rceil, \\ \tilde{N}^{\max} = \lceil \frac{3N}{2G} \rceil. \end{cases} \tag{6}$$

Eq. (4) is the cardinality constraint. Eq. (5) corresponds to an indicator function showing whether asset j belongs to cluster g. Eq. (6) specifies the minimum and maximum asset numbers in a cluster. The above optimization problem is hard to solve by using traditional optimization methods. Brucker [1978] pointed out that clustering problems turn out to be non-deterministic polynomial-time hard (NP-hard) when the cluster number G becomes higher.

3 Asset Allocation Methods

3.1 Weight Constraints

As with traditional asset allocation problems, the budget constraint must be met while using the proposed clustering asset allocation. The sum of cluster member weights in a cluster should be equal to 1, and likewise the sum of cluster portfolio weights. As the current asset allocation problem does not consider short sales of assets, the budget constraints can be written as

$$\sum_{g=1}^{G} w_g = 1, \quad \text{and} \quad \sum_{i} w_{g,i} = 1 \quad \text{for} \quad i \in \mathcal{C}_g, \tag{7}$$

with

$$w_g \geq 0, \quad w_{g,i} \begin{cases} \geq 0 & i \in \mathcal{C}_g, \quad 1 \leq g \leq G, \\ = 0 & \text{otherwise.} \end{cases} \tag{8}$$

w_g denotes the weight of the g-th cluster portfolio, and $w_{g,i}$ represents the weight of the i-th cluster member in the g-th cluster $\mathcal{C}_g$. The actual weight of asset j is denoted by w_j, which is the product of its corresponding cluster member's weight $w_{g,i}$ in the cluster and the cluster portfolio weight w_g

$$w_j = w_g \cdot w_{g,i}, \quad i \in \mathcal{C}_g, \quad 1 \leq g \leq G, \quad 1 \leq j \leq N. \tag{9}$$

3.2 Equal Weight Allocation

In the equal weight (EW) allocation, the cluster portfolio weights are related to the cluster number G, and the cluster member weights are dependent on the cluster size. Although the cluster number G is manually assigned, the final weight of an asset depends on the clustering structure since the weight of a cluster member is actually decided by the clustering structure. In other words, the cluster portfolio weight w_g is given by 1 over the cluster number G, and the cluster member weight $w_{g,i}$ in the subset $\mathcal{C}_g$ can be calculated by taking 1 over the number of members in the cluster:

$$w_g = \frac{1}{G}, \tag{10}$$

$$w_{g,i} = \frac{1}{\sharp\{\mathcal{C}_g\}}, \quad i \in \mathcal{C}_g, \quad 1 \leq g \leq G. \tag{11}$$

3.3 Minimum Variance Allocation

In the literature, the minimum variance portfolio (MVP) is regarded as the safest portfolio due to its minimum variance at the Markowitz efficient frontier (see Markowitz [1952]). Quadratic and concave programming are usually applied to solve optimization problems which have quadratic objective functions with linear equality and inequality constraints. The objective function of traditional Markowitz mean-variance allocation can be formulated as

$$\max_{\mathbf{w}} \zeta \mathbf{r}'\mathbf{w} - (1-\zeta)\mathbf{w}'\Sigma\mathbf{w}, \tag{12}$$

where $\mathbf{w}$ denotes the weight vector for either the cluster members or the cluster portfolios, and Σ represents the variance-covariance matrix of either the cluster portfolios or the cluster members. $\mathbf{r}$ is referred to as the expected return vector of assets, and ζ is a risk aversion factor (or a weighting difference factor) with a value of 0 for a minimum variance allocation.

4 Optimization Methods

4.1 Differential Evolution

Heuristic methods have been documented in the literature as they provide solutions to tackle complex constrained optimization problems in economics (see Gilli and Winker [2008]). Two population-based evolutionary methods (Differential Evolution and Genetic Algorithm) are used to tackle the clustering problem for the purpose of the SR maximization. The first evolutionary method used to tackle the optimization problem is Differential Evolution (DE) proposed by Storn and Price [1997], which is a population-based local search heuristic method for solving optimization problems with a continuous solution space. This approach generates new solutions by using mutation and crossover (i.e. a process of linear combination with three current solutions). More specifically, let $\mathfrak{P}$ denote the number of solutions in each generation (i.e. the population size), then, for each current solution $\dot{\mathbf{s}}_{\mathfrak{p}}$, a new solution $\dot{\mathbf{s}}_{\mathfrak{c}}$ is generated by linearly combining the solution vectors of three members ($\mathfrak{p}_1 \neq \mathfrak{p}_2 \neq \mathfrak{p}_3 \neq \mathfrak{p}$) of the current population

$$\dot{s}_{\mathfrak{c}}[i] := \begin{cases} \dot{s}_{\mathfrak{p}_1}[i] + (K + z_1[i]) \cdot (\dot{s}_{\mathfrak{p}_2}[i] - \dot{s}_{\mathfrak{p}_3}[i] + z_2[i]) & \text{with probability } \mathfrak{z}_1, \\ \dot{s}_{\mathfrak{p}}[i] & \text{otherwise,} \end{cases} \tag{13}$$

where $\mathfrak{z}_1$ is the crossover probability.

The algorithm is the 'Dither' and 'Jitter' version of the standard DE (see Price et al. [1998]) which involves adding normally distributed random numbers to the

Algorithm 1 Differential Evolution

Randomly initialize population of vectors $\dot{\mathbf{s}}_{\mathfrak{p}}$, $\mathfrak{p}$= 1,...,$\mathfrak{P}$
while the iteration number is not met **do**
  for all current solutions $\dot{\mathbf{s}}_{\mathfrak{p}}$, $\mathfrak{p}$=1,...,$\mathfrak{P}$ **do**
    Randomly pick three different solutions, i.e. $\mathfrak{p}_1 \neq \mathfrak{p}_2 \neq \mathfrak{p}_3 \neq \mathfrak{p}$
    $\dot{s}_{\mathfrak{c}}[i] \leftarrow \dot{s}_{\mathfrak{p}_1}[i] + (K + z_1[i])(\dot{s}_{\mathfrak{p}_2}[i] - \dot{s}_{\mathfrak{p}_3}[i] + z_2[i])$ with probability $\mathfrak{z}_1$, or $\dot{s}_{\mathfrak{c}}[i] \leftarrow \dot{s}_{\mathfrak{p}}[i]$ otherwise
    Compute the fitness value of $\dot{\mathbf{s}}_{\mathfrak{p}}$
  end for
  for the current solution $\dot{\mathbf{s}}_{\mathfrak{p}}$, $\mathfrak{p}$ = 1,...,$\mathfrak{P}$ **do**
    if Fitness($\dot{\mathbf{s}}_{\mathfrak{c}}$) > Fitness($\dot{\mathbf{s}}_{\mathfrak{p}}$) **then** $\dot{\mathbf{s}}_{\mathfrak{p}} \leftarrow \dot{\mathbf{s}}_{\mathfrak{c}}$ **end if**
  end for
end while

weighting factor K and the difference of the two solution vectors, in order to increase the pool of potential trial vectors and minimize the risk of getting stuck in local optima. Vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ represent the two random noises which are added to the weighting factor and the vector difference respectively. The two vectors are actually generated by using random numbers, which are zeros at the probabilities of $\mathfrak{z}_2$ and $\mathfrak{z}_3$ respectively or which follow normal distributions $\mathcal{N}(0, \mathfrak{d}_1^2)$ and $\mathcal{N}(0, \mathfrak{d}_2^2)$ otherwise. After the above process, DE updates the population on the basis of Charles Darwin's concept of natural selection. Thus, if the fitness value of $\dot{\mathbf{s}}_{\mathfrak{c}}$ is higher than that of $\dot{\mathbf{s}}_{\mathfrak{p}}$, then $\dot{\mathbf{s}}_{\mathfrak{p}}$ is replaced by $\dot{\mathbf{s}}_{\mathfrak{c}}$ and the updated $\dot{\mathbf{s}}_{\mathfrak{p}}$ exists in the current population; otherwise $\dot{\mathbf{s}}_{\mathfrak{p}}$ survives in the current population. The fitness value is the SR value defined in Eq. (1). To meet the integer constraint, the solutions are rounded to integers. Additionally, a punishment scheme has been utilized to impair the fitness of solutions which violate the cardinality constraints. The pseudo code in Algorithm 1 describes the DE procedure.

4.2 Genetic Algorithm

Genetic Algorithms (GAs) were first used by Holland [1975]. These algorithms have been proposed and successfully applied in solving clustering problems in computer science (see Krovi [1992] and Hruschka et al. [2009]). GAs are search algorithms which adopt selection mechanics based on natural selection and genetics. Each gene of a chromosome (or an individual solution) represents the cluster an asset should belong to. GAs are usually inspired by biological evolution that is normally developed in three steps: natural selection, crossover and mutation. In the selection process, the algorithm chooses chromosomes from the current population for mutation. In other words, the process applies natural selection (i.e. the Darwinian survival rule of the fittest) to select the string representatives. Roulette

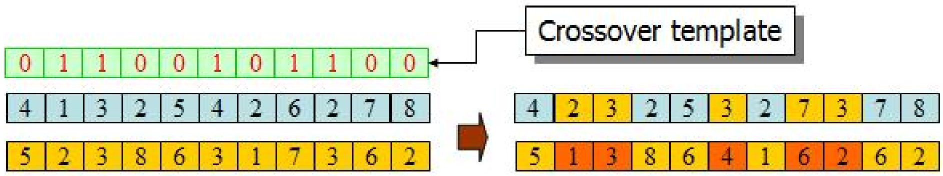

Figure 2: The Uniform Crossover Scheme

wheel selection (see Holland [1975]) is usually employed to implement the proportional selection in GAs. A suitable fitness evaluation could be the SR value defined in Eq. (1). However, since the SR value can be negative (this is not desirable for the canonical roulette wheel selection method), the actual fitness function used is the one given by Eq. (14), where 1 is added to the SR value in order to guarantee a positive fitness value in general cases

$$\widetilde{SR} = 1 + \frac{r_P - r_f}{\sigma_P}. \tag{14}$$

The fitness values are used to associate a probability of selection with each individual chromosome. Let $\widetilde{SR}_{\mathfrak{p}}$ denote the fitness of the $\mathfrak{p}$-th individual chromosome in the population, then its probability of being selected is decided by

$$\mathfrak{z}_{\mathfrak{p}} = \frac{\widetilde{SR}_{\mathfrak{p}}}{\sum_{n=1}^{\mathfrak{P}} \widetilde{SR}_n}, \quad \mathfrak{p} \in \mathfrak{P}, \tag{15}$$

where $\mathfrak{P}$ is the population size, i.e. the number of chromosomes in the population.

After the selection, a probabilistic process (crossover), which exchanges information between two parent chromosomes is performed to produce child chromosomes. There are several crossover schemes available in the literature, such as the single-point, the multi-point and the uniform crossover. The uniform crossover scheme, which exchanges corresponding genes in the parent chromosomes with probability $\mathfrak{z}_c$, is adopted. After the crossover, genes in each chromosome undergo a mutation process which introduces gene variation into the population. For a binary representation of a gene, the mutation can be implemented simply by flipping the binary value of genes. In this clustering optimization problem, the genes are mutated with probability $\mathfrak{z}_m$ by replacing the original genes with uniformly distributed random integer numbers between 1 and G. If a solution does not meet the cardinality constraints, the fitness value of the solution is impaired by using a punishment mechanism. After mutation, each parent solution in the current

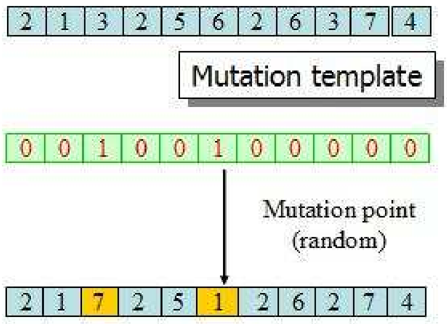

Figure 3: The Mutation Scheme

population is replaced by its offspring if the latter has a higher fitness value. The above processes are repeated until a fixed number of generations reached. The GA algorithm is described in Algorithm 2.

Algorithm 2 Genetic Algorithm for Clustering Optimization

Randomly initialize population of vectors $\dot{\mathbf{s}}_{\mathfrak{p}}$, $\mathfrak{p}$= 1...$\mathfrak{P}$
Evaluate the initial population
while a fixed number of generations is not met **do**
    Select parent chromosomes based on the roulette wheel selection with probability $\mathfrak{z}_{\mathfrak{p}} = \frac{\widetilde{SR}_{\mathfrak{p}}}{\sum_{n=1}^{\mathfrak{P}} \widetilde{SR}_n}$
    Generate $\dot{\mathbf{s}}_{\mathfrak{c}}$ by applying the uniform crossover and the mutation scheme at probabilities $\mathfrak{z}_c$ and $\mathfrak{z}_m$
    Evaluate the offspring $\dot{\mathbf{s}}_{\mathfrak{c}}$
    if Fitness($\dot{\mathbf{s}}_{\mathfrak{c}}$) $>$ Fitness($\dot{\mathbf{s}}_{\mathfrak{p}}$)
    then $\dot{\mathbf{s}}_{\mathfrak{p}} \leftarrow \dot{\mathbf{s}}_{\mathfrak{c}}$
end while

---

5 The Experiments

5.1 Data and Algorithm Settings

The in-sample and out-of-sample studies involve daily log-return of 85 stocks selected from the FTSE market for the period January 2005 to December 2009. The

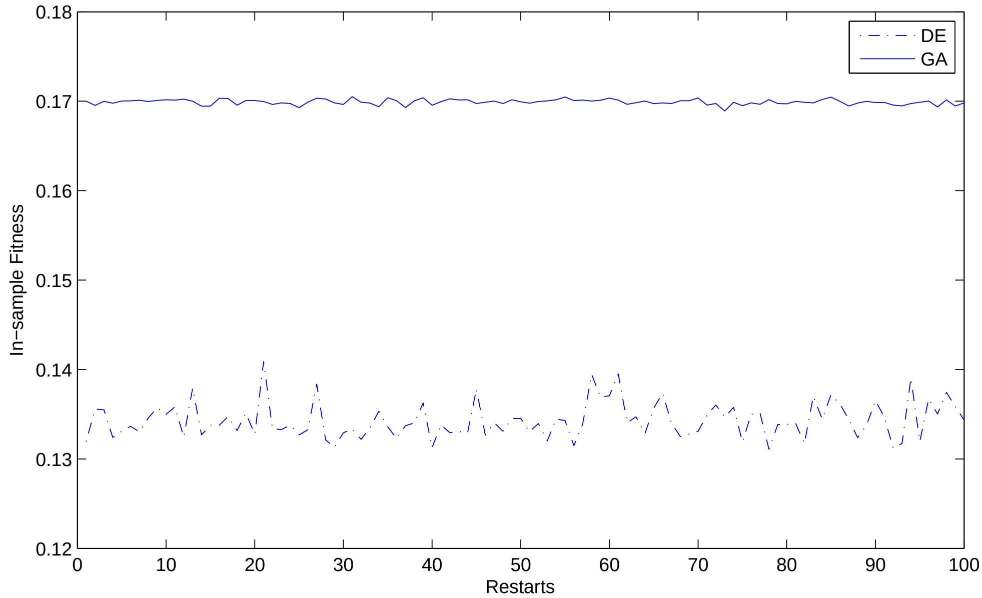

Figure 4: Sharpe Ratio from Independent Restarts

proposed clustering asset allocation is applied to construct portfolios based on one year's data, and then the portfolios are held in the consecutive year for out-of-sample studies. As a consequence, there are four in-sample and four out-of-sample periods. The adjusted daily 1-year LIBOR rate is considered as the risk-free rate in computing the SR.

To assess the performance of GA and DE in tackling the clustering problem, fitness values from 100 independent restarts of the two algorithms are compared with each other by using the same number of generations and the same population size. DE has been used to explore the clustering impact on portfolio performance by Zhang and Maringer [2009]; the parameters of DE used in this paper are same as the DE settings in that paper. Figure 4 shows the fitness values, which reveals GA outperforming DE in terms of higher fitness values from the independent restarts. It is reasonable since DE was originally designed for optimization problems with a continuous solution space. While using DE to solve the clustering problem, the solutions are rounded to integers to meet the integer constraint which may slow down the evolutionary procedure. In contrast to DE, GA directly uses integers as chromosome representations. As a result, GA is able to obtain higher fitness values and converge faster than DE by using the same generation number and the same population size.

Based on preliminary tests, the following parameters of the GA were found to be suitable for solving the optimization problem. The settings were specified as follows: each chromosome had 85 genes representing the 85 stocks; the population size was set at 340; the number of generations was set at 8,000; the crossover probability was 0.5; and the mutation probability was 0.2.

5.2 In-Sample Study

5.2.1 Sharpe Ratio from the Bootstrap Method

To reveal the clustering impact on SR, the in-sample SR distributions were constructed on the basis of simulated asset returns with the optimized weights from the optimal clustering structure. The simulated asset returns were bootstrapped from the in-sample asset returns, and the portfolio weights were distributed by using the EW strategy and the MVP allocation based on the in-sample asset returns. The daily simulated asset returns of each in-sample period (2005, 2006, 2007 and 2008) were generated by using the historical asset returns of the corresponding year with a length of 260 and a bootstrapped iteration number of 2,000.

Ideally, the first and the third moments (i.e. the mean value and the skewness) of the portfolio SR distribution under the clustering asset allocation structure should be higher than the moments of the SR distribution without the clustering effect; the second and the fourth moments (i.e. the standard deviation and the kurtosis) of the SR distributions under the clustering impact should stick close to the two moments in the non-clustered case.

Table 1 shows the descriptive statistics of the bootstrapped SR distributions on the basis of portfolio weights from the EW strategy. It is found that the mean and the skewness (SK) of the SR distribution in the clustered cases are higher than the two moments in the non-clustered case. Moreover, the standard deviation and the kurtosis (KU) of the SR distributions are found to be approximately the same as the two moments without the clustering effect. Table 2 provides the statistics of the SR distributions when using the MVP allocation. In most cases, the mean values and the skewness under the clustering impact, are higher than those in the non-clustered case. Furthermore, the standard deviation and the kurtosis after applying the clustering scheme, remain almost the same as the two moments of the SR distribution in the non-clustered case.

Table 1: Descriptive Statistics of Bootstrapped In-Sample SR with EW

	2005		2006		2007		2008	
	MEAN	STDV	MEAN	STDV	MEAN	STDV	MEAN	STDV
non-clustered	0.1317	0.0670	0.0571	0.0627	-0.0058	0.0641	-0.0752	0.0630
$G = 3$	0.1773	0.0681	0.0873	0.0631	0.0285	0.0645	-0.0587	0.0629
$G = 5$	0.1805	0.0680	0.0897	0.0632	0.0302	0.0645	-0.0580	0.0629
$G = 7$	0.1814	0.0678	0.0900	0.0633	0.0294	0.0645	-0.0582	0.0628
$G = 9$	0.1847	0.0682	0.0924	0.0632	0.0309	0.0644	-0.0571	0.0627
	SK	KU	SK	KU	SK	KU	SK	KU
non-clustered	0.0397	2.9414	0.1436	2.9354	0.0130	2.8143	-0.0363	2.9297
$G = 3$	0.0432	2.9260	0.1563	2.9864	0.0262	2.8350	-0.0265	2.9503
$G = 5$	0.0469	2.9096	0.1612	2.9923	0.0294	2.8383	-0.0242	2.9563
$G = 7$	0.0480	2.9186	0.1626	2.9852	0.0317	2.8342	-0.0238	2.9559
$G = 9$	0.0512	2.9247	0.1568	2.9838	0.0268	2.8279	-0.0216	2.9570

Table 2: Descriptive Statistics of Bootstrapped In-Sample SR with MVP

	2005		2006		2007		2008	
	MEAN	STDV	MEAN	STDV	MEAN	STDV	MEAN	STDV
non-clustered	0.0860	0.0651	0.0639	0.0621	0.0175	0.0643	-0.0216	0.0624
$G = 3$	0.1419	0.0664	0.0453	0.0624	0.0451	0.0643	-0.0095	0.0628
$G = 5$	0.0907	0.0655	0.0583	0.0622	0.0142	0.0642	-0.0028	0.0629
$G = 7$	0.1820	0.0668	0.1346	0.0637	0.0613	0.0642	0.0006	0.0629
$G = 9$	0.1865	0.0669	0.1422	0.0637	0.0629	0.0645	0.0008	0.0629
	SK	KU	SK	KU	SK	KU	SK	KU
non-clustered	0.0647	3.0133	0.0510	3.0264	-0.0165	2.8873	0.0198	3.0360
$G = 3$	0.1174	3.0225	0.0566	3.0703	-0.0166	2.8549	0.0215	3.0679
$G = 5$	0.0369	3.0159	0.0806	3.0565	-0.0124	2.8872	0.0310	3.0745
$G = 7$	0.0989	2.9719	0.1086	3.0452	-0.0021	2.8508	0.0261	3.0807
$G = 9$	0.0860	2.9855	0.0967	3.0383	0.0058	2.8647	0.0266	3.0996

5.2.2 Sharpe Ratio from the Single Index Market Model

A classic market model was adopted in the second in-sample experiment in order to provide asset return simulations for the clustering effect study. As the single index market (SIM) model has been documented and discussed extensively in the literature, it might be helpful to further explore the clustering effect in this paper. The SIM model can be considered as a single-factor regression model of the asset returns depending on the returns of a market portfolio

$$r_j = \alpha_j + \beta_j \cdot r_M + \epsilon_j. \tag{16}$$

By resorting to the least squares estimation, one can estimate the intercept α_j and the slope β_j for asset j based on the asset return r_j and the market portfolio return r_M. The FTSE index daily log-return was considered as the market portfolio return. Since it is usually assumed that the residuals from the regression follow normal distributions, simulations of daily asset return $\hat{r}_j$ can be generated by using the SIM model after estimating the intercept $\hat{\alpha}_j$, the slope $\hat{\beta}_j$, the mean and the standard deviation of the residual ϵ_j. In other words, for each in-sample period (2005, 2006, 2007 and 2008), the asset returns and the FTSE return were used to estimate the model parameters and the residual parameters for the four periods. The simulated asset returns were then generated by using the SIM model with the estimated model parameters and the artificially generated random noises. Each artificially generated asset return series comprises an observation number of 260 and a simulation path of 2,000 in each period.

Table 3 provides the statistics of the SR distributions which are computed by using the simulated returns from the SIM model and the portfolio weights from the EW allocation. As the table shows, the mean value and skewness of the SR distributions from using the clustering asset allocation are higher than the two moments in the non-clustered case. The standard deviation and the kurtosis, after applying the clustering scheme, are slightly higher than the two moments in the absence of clustering impact. The above findings are consistent with those observed in the previous in-sample bootstrap study. Table 4 reports the moment statistics of the SR distributions when applying the MVP allocation. Generally, these statistics agree with the clustering impact observed before, i.e. the mean value are increased by using the clustering scheme while the standard deviation and the kurtosis are kept almost the same as the two moments in the non-clustered case. However, one should note that some skewness ratios are not improved under the clustering impact (e.g. for the ratios in 2005 and 2008).

Table 3: Descriptive Statistics of In-Sample SR with SIM and EW

	2005		2006		2007		2008	
	MEAN	STDV	MEAN	STDV	MEAN	STDV	MEAN	STDV
non-clustered	-0.0576	0.0036	-0.0613	0.0033	-0.0698	0.0040	-0.0750	0.0079
$G = 3$	-0.0455	0.0044	-0.0506	0.0039	-0.0537	0.0048	-0.0582	0.0083
$G = 5$	-0.0447	0.0045	-0.0496	0.0039	-0.0529	0.0048	-0.0575	0.0084
$G = 7$	-0.0443	0.0044	-0.0496	0.0040	-0.0533	0.0047	-0.0578	0.0085
$G = 9$	-0.0433	0.0046	-0.0486	0.0041	-0.0526	0.0049	-0.0568	0.0087
	SK	KU	SK	KU	SK	KU	SK	KU
non-clustered	-0.0324	2.8525	0.0224	3.0778	0.0137	2.9111	-0.0071	2.7766
$G = 3$	0.0168	2.9052	0.0615	3.2635	0.0169	2.9069	0.0600	2.7501
$G = 5$	0.0365	2.8824	0.0579	3.3206	0.0107	2.8708	0.0622	2.7550
$G = 7$	0.0071	2.9400	0.0490	3.3759	0.0415	2.8652	0.0349	2.6647
$G = 9$	0.0523	2.9454	0.0887	3.1736	0.0449	2.8167	0.0162	2.7722

Table 4: Descriptive Statistics of In-Sample SR with SIM and MVP

	2005		2006		2007		2008	
	MEAN	STDV	MEAN	STDV	MEAN	STDV	MEAN	STDV
non-clustered	-0.0659	0.0092	-0.0546	0.0116	-0.0569	0.0128	-0.0233	0.0275
$G = 3$	-0.0523	0.0079	-0.0622	0.0107	-0.0437	0.0112	-0.0107	0.0301
$G = 5$	-0.0664	0.0065	-0.0581	0.0095	-0.0589	0.0109	-0.0037	0.0270
$G = 7$	-0.0416	0.0085	-0.0269	0.0107	-0.0352	0.0118	-0.0007	0.0291
$G = 9$	-0.0401	0.0084	-0.0229	0.0113	-0.0336	0.0126	-0.0005	0.0287
	SK	KU	SK	KU	SK	KU	SK	KU
non-clustered	0.0393	2.9252	-0.1472	3.2116	-0.0168	2.9697	0.0037	2.7621
$G = 3$	-0.0101	3.0362	-0.0942	3.1220	-0.0801	2.9141	-0.0307	2.8588
$G = 5$	0.0321	2.9459	-0.0617	3.0740	0.0426	3.0439	-0.0297	2.9463
$G = 7$	-0.0396	2.9370	-0.0217	2.8195	-0.0259	2.8228	-0.0523	2.9624
$G = 9$	-0.0341	2.9162	-0.0624	3.0335	0.0499	2.9149	-0.0466	2.9951

5.3 Out-of-Sample Study

5.3.1 Sharpe Ratio from the Bootstrap Method

If the clustering design were to improve the risk-adjusted reward measure during the in-sample period, one would expect a corresponding impact for the out-of-sample period. Thus, it is necessary to study the clustering impact on the out-of-sample SR distributions which are constructed by using the same approaches as in the in-sample studies. The out-of-sample period was considered as the consecutive year of each in-sample period. In other words, each in-sample period (2005, 2006, 2007 and 2008) has a corresponding out-of-sample period (2006, 2007, 2008 and 2009) respectively. The portfolio weights which were optimized by using the clustering asset allocation scheme, with different cluster numbers in the in-sample periods, were held in the corresponding out-of-sample period. The simulated asset returns series were generated by using the bootstrap approach on the basis of the real asset returns in each out-of-sample period with a length of 260 observations and 2,000 iterations.

Table 5 provides the moment statistics of the bootstrapped SR distributions which were constructed by using the portfolio weights from adopting the EW strategy and the bootstrapped out-of-sample asset returns. From the table, it can be known that the mean values and the skewness under the clustering impact are higher than the two moments in the non-clustered case, and that the standard deviation and the kurtosis remain approximately the same as those in the non-clustered case. The findings are consistent with those observed in the in-sample studies. Table 6 reports the moment statistics of the out-of-sample bootstrapped SR distributions based on the MVP weights. Although it is hard to observe a clear clustering impact on the second, third and fourth moments of the distributions (standard deviation, skewness and kurtosis), it can be seen that most of the mean values in the clustered cases are still higher than that in the non-clustered case, except for 2008, the U.S. financial crisis year.

5.3.2 Sharpe Ratio from the Single Index Market Model

The SIM model, which was introduced in the in-sample study, has been employed again to simulate asset returns for the out-of-sample study. The simulated returns in each out-of-sample period were generated by using the SIM model after estimating the model parameters and the residual parameters based on the asset returns and the FTSE index return in each out-of-sample period. Each artificially generated asset return series in each out-of-sample period contains 260 observations and 2,000 simulations.

Table 7 reports the moments of the SR distributions based on the weights from the EW strategy and the simulated asset returns from the SIM model. It is clear

Table 5: Descriptive Statistics of Bootstrapped Out-of-Sample SR with EW

	2006		2007		2008		2009	
	MEAN	STDV	MEAN	STDV	MEAN	STDV	MEAN	STDV
non-clustered	0.0583	0.0628	-0.0049	0.0626	-0.0811	0.0636	0.0523	0.0647
$G = 3$	0.0614	0.0632	-0.0013	0.0627	-0.0761	0.0637	0.0577	0.0648
$G = 5$	0.0590	0.0631	-0.0015	0.0627	-0.0771	0.0637	0.0572	0.0647
$G = 7$	0.0575	0.0632	-0.0012	0.0627	-0.0765	0.0637	0.0578	0.0647
$G = 9$	0.0579	0.0632	-0.0013	0.0628	-0.0755	0.0638	0.0597	0.0648
	SK	KU	SK	KU	SK	KU	SK	KU
non-clustered	0.1449	2.9360	0.1121	3.1714	-0.0882	3.0633	0.1241	2.9880
$G = 3$	0.1560	2.9011	0.1345	3.1834	-0.0795	3.0796	0.1250	2.9977
$G = 5$	0.1606	2.8994	0.1342	3.1869	-0.0776	3.0795	0.1263	3.0036
$G = 7$	0.1598	2.8836	0.1357	3.1763	-0.0762	3.0726	0.1291	3.0073
$G = 9$	0.1557	2.8870	0.1332	3.1930	-0.0783	3.0888	0.1273	3.0053

Table 6: Descriptive Statistics of Bootstrapped Out-of-Sample SR with MVP

	2006		2007		2008		2009	
	MEAN	STDV	MEAN	STDV	MEAN	STDV	MEAN	STDV
non-clustered	0.0542	0.0621	0.0164	0.0630	-0.0459	0.0646	0.0385	0.0647
$G = 3$	0.0575	0.0626	0.0186	0.0629	-0.0489	0.0646	0.0537	0.0637
$G = 5$	0.0594	0.0622	0.0018	0.0632	-0.0472	0.0646	0.0546	0.0635
$G = 7$	0.0545	0.0625	0.0203	0.0634	-0.0504	0.0645	0.0447	0.0639
$G = 9$	0.0579	0.0627	0.0221	0.0634	-0.0527	0.0644	0.0420	0.0640
	SK	KU	SK	KU	SK	KU	SK	KU
non-clustered	0.1353	2.9979	0.1087	3.1653	-0.0354	3.0608	0.0840	3.0271
$G = 3$	0.1412	2.9903	0.1117	3.1476	-0.0268	3.0626	0.0082	2.9122
$G = 5$	0.1223	3.0044	0.0930	3.1546	-0.0403	3.0499	0.0657	2.9605
$G = 7$	0.1372	2.9897	0.1584	3.1117	-0.0318	3.0836	0.0499	2.9248
$G = 9$	0.1579	3.0032	0.1533	3.1659	-0.0376	3.0659	0.0552	2.9157

that the mean values and the skewness of the SR distributions under the clustering impact are higher than the moments of the SR distributions without clustering. The standard deviation and the kurtosis of the SR distributions in the clustered cases are slightly higher than the moments without the clustering impact.

Table 8 shows the SR distribution statistics obtained from adopting the MVP weights based on the simulated returns. Although the impact of the clustering scheme on the SR distribution is not clear as those observed from the in-sample studies, it is still found that the mean values of the SR distribution in 2007 and 2009 under the clustering impact are higher than those without clustering. It should be noted that the mean values from the clustering asset allocation do not outperform the non-clustered means in 2006 and 2008, and the standard deviations under the clustering impact are correspondingly smaller than the non-clustered standard deviations in the two years (which is also observed in the case of 2008 from Table 6). This may imply, by utilizing the clustering design, statistical inference approaches (such as the bootstrap and the simulation method) are able to provide better confidence intervals for the population means than the approaches in the absence of the clustering design when the market is experiencing a structural break. In such cases, the mean values as well as the standard deviations may decrease after using the clustering scheme. Regarding the other two moments, it seems that the clustering scheme does not have any significant impact on the skewness and the kurtosis of the SR distribution.

6 Conclusion

This paper presents an approach which utilizes a clustering technique with traditional asset allocation methods to improve the portfolio Sharpe ratio. The clustering approach in question is different from traditional clustering techniques with regards to the maximization of the in-sample SR being used as the clustering criterion. The portfolio weights are determined by the weights of two parts (i.e. the cluster portfolios and the cluster members), which are computed by adopting the EW strategy and the Markowitz MVP allocation in this paper. The comparative study on the performance of DE against GA shows that GA is more apt at finding the optimal cluster structures.

To explore the clustering impact on the portfolio SR distribution, the traditional bootstrap approach and the single index model are used to provide simulations of asset returns for the in-sample and out-of-sample studies. As the in-sample experiment results show, the SR distributions become better after applying the clustering asset allocation: higher mean and skewness values are able to achieve, and changes in the standard deviation and kurtosis are small with the equal weight investment strategy. With the minimum variance allocation, the mean value of the

Table 7: Descriptive Statistics of Out-of-Sample SR with SIM and EW

	2006		2007		2008		2009	
	MEAN	STDV	MEAN	STDV	MEAN	STDV	MEAN	STDV
non-clustered	0.0603	0.0053	0.0357	0.0060	0.0299	0.0125	0.0517	0.0113
$G = 3$	0.0627	0.0061	0.0393	0.0066	0.0380	0.0136	0.0567	0.0113
$G = 5$	0.0614	0.0062	0.0384	0.0069	0.0363	0.0137	0.0561	0.0115
$G = 7$	0.0606	0.0061	0.0405	0.0070	0.0374	0.0136	0.0567	0.0115
$G = 9$	0.0609	0.0062	0.0397	0.0070	0.0390	0.0142	0.0586	0.0117
	SK	KU	SK	KU	SK	KU	SK	KU
non-clustered	0.0855	2.8542	-0.1609	2.7678	0.0142	3.1440	0.0592	3.0971
$G = 3$	0.0847	2.9164	-0.1173	2.8901	0.0144	3.1618	0.0865	3.1521
$G = 5$	0.0861	2.9054	-0.1084	2.8639	0.0145	3.1543	0.0643	3.0760
$G = 7$	0.0855	2.9168	-0.1371	2.8516	0.0146	3.2785	0.0620	3.0912
$G = 9$	0.0877	2.9471	-0.1407	2.9198	0.0140	3.2035	0.0824	3.1196

Table 8: Descriptive Statistics of Out-of-Sample SR with SIM and MVP

	2006		2007		2008		2009	
	MEAN	STDV	MEAN	STDV	MEAN	STDV	MEAN	STDV
non-clustered	0.0659	0.0157	0.0593	0.0168	0.0723	0.0298	0.0390	0.0379
$G = 3$	0.0615	0.0125	0.0615	0.0169	0.0695	0.0272	0.0558	0.0443
$G = 5$	0.0622	0.0125	0.0485	0.0152	0.0733	0.0268	0.0570	0.0416
$G = 7$	0.0599	0.0126	0.0626	0.0173	0.0667	0.0289	0.0471	0.0434
$G = 9$	0.0617	0.0128	0.0642	0.0173	0.0614	0.0301	0.0439	0.0428
	SK	KU	SK	KU	SK	KU	SK	KU
non-clustered	0.0745	2.8871	-0.0393	2.7884	0.0921	3.2321	0.0087	2.9595
$G = 3$	0.0034	2.8509	-0.0365	2.8268	-0.0162	3.1059	-0.0781	3.0963
$G = 5$	0.0705	2.9044	-0.0406	2.8870	0.1602	3.1859	-0.0647	3.0556
$G = 7$	-0.0477	2.7892	0.0042	3.0750	-0.1027	3.1043	-0.0421	3.1802
$G = 9$	-0.0526	2.7041	-0.0012	3.1214	-0.1057	2.8430	-0.0311	3.1605

SR distributions is increased whereas the other three moments of the distributions are not significantly improved by using the clustering scheme. The results from the out-of-sample experiment further support the findings from the in-sample studies. More importantly, this paper reveals that by using the clustering design with existing asset allocations, it is possible to obtain a portfolio with more beneficial features by using a proper clustering criterion. For instance, the portfolio SR is improved after adopting the SR maximization as the clustering criterion. Portfolio managers may therefore apply the proposed asset allocation scheme by changing the clustering criterion to obtain portfolios with desired properties which better suit the preferences and demands of their clients.

References

Brucker, P. (1978). On the complexity of clustering problems. In Henn, R., Korte, B., and Oettli, W., editors, *Optimization and Operations Research*, volume 157, pages 45–54. Springer, Berlin.

Gilli, M. and Winker, P. (2008). A review of heuristic optimization methods in econometrics. *Swiss Finance Institute Research Paper, 08–12.*

Harris, R. D. and Yilmaz, F. (2007). Estimation of the conditional variance – covariance matrix of returns using the intraday range. Xfi - Centre for Finance and Investment, University of Exeter.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems: An introductory analysis with applications to biology, control, and artificial intelligence.* University of Michigan Press, Ann Arbor.

Hruschka, E., Campello, R., Freitas, A., and de Carvalho, A. (2009). A survey of evolutionary algorithms for clustering. In *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, volume 39, pages 133–155. Institute of Electrical and Electronics Engineers Inc.

Krovi, R. (1992). Genetic algorithms for clustering: A preliminary investigation. In Milutinovic, V., Shriver, B. D., Numaker, J. F. J., and Sprague, R. H. J., editors, *Proceedings of the 25 Hawaii International Conference on System Sciences*, volume 4, pages 540–544. IEEE Computer Society Press.

Lisi, F. and Corazza, M. (2008). Clustering financial data for mutual fund management. In Perna, C. and Sibillo, M., editors, *Mathematical and Statistical Methods in Insurance and Finance*, pages 157–164. Springer, Milan.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7:77–99.

Pattarin, F., Paterlini, S., and Minerva, T. (2004). Clustering financial time series: An application to mutual funds style analysis. *Computational Statistics and Data Analysis*, 47:353–372.

Price, K., Storn, R., and Lampinen, J. (1998). *Differential Evolution: A practical approach to global optimization.* Springer, Berlin.

Storn, R. and Price, K. (1997). Differential Evolution – A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11:341–359.

Windcliff, H. and Boyle, P. (2004). The 1/n pension investment puzzle. *North American Actuarial Journal*, 8:32–45.

Zhang, J. and Maringer, D. (2009). Improving Sharpe ratios and stability of portfolios by using a clustering technique. In *Proceedings of the 2009 International Conference of Computational Intelligence and Intelligent Systems, Lecture Notes in Engineering and Computer Science, WCE*, pages 1–6.