Empirical Economics (2020) 59:2423–2442
https://doi.org/10.1007/s00181-019-01802-5

Development of an efficient cluster-based portfolio optimization model under realistic market conditions

Mahdi Massahi[1] · Masoud Mahootchi[1] · Alireza Arshadi Khamseh[2]

Received: 30 August 2016 / Accepted: 29 July 2019 / Published online: 21 January 2020

Abstract

Modern portfolio theory introduced by Markowitz in 1952 is the most popular portfolio optimization framework established based on the trade-off between risk and return as an operation research model. The main shortcoming of applying Markowitz portfolio optimization in practice is that the obtained optimal weights are really sensitive to the embedded uncertainty in return series of stocks. In this paper, it is demonstrated how using a new methodology of time series clustering as a remedy can lead to a more robust and accurate portfolio in terms of the gap between mean variance efficient frontier obtained from the optimization model and the one observed in reality. In this regard, two similarity measures, the autocorrelation coefficients and the weighted dynamic time warping, are used in an innovative way to construct the desired portfolio optimization model. Moreover, the effectiveness of proposed approach is investigated in two different market conditions: semi-realistic and full-realistic. In the first one, it is assumed that the forecasted and realized stocks mean returns are the same; however, these returns are not necessarily equal in the second market conditions. Finally, a database of stock prices from the literature is utilized to show the robustness and accuracy of the proposed approach in empirical results in comparison with applied similarity measures in previous researches.

Keywords Portfolio optimization · Realized efficient frontier · Time series clustering · Weighted dynamic time warping · Autocorrelation coefficient · Realistic market condition

✉ Masoud Mahootchi
mmahootchi@aut.ac.ir; mmahootchi@gmail.com

Mahdi Massahi
m.massahi@aut.ac.ir

Alireza Arshadi Khamseh
ar_arshadi@yahoo.com

[1] Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, 424 Hafez Ave., Tehran, Iran

[2] Department of Industrial Engineering, Kharazmi University, 43 Shahid Mofateh Ave., Tehran, Iran

JEL Classification G11 · C61 · C63 · C38

1 Introduction

Nowadays, portfolio management and finding an optimal portfolio are common issues in finance. The modern portfolio theory was firstly reviewed in the work of Markowitz (1959) and Sharpe (1964) who were awarded the Nobel Prize in Economics in 1990. Since 1990, many authors have investigated how to find the optimal portfolio and a wide range of methods for portfolio selection and optimization have been used (see, for example, Ehrgott et al. 2004; Lin and Liu 2008; Best 2010; Vaclavik and Jablonsky 2012; Cong and Oosterlee 2016). Financial experts believe that diversification of investment is vital to create an efficient portfolio. Tola et al. (2008) indicated that, to make a diversified portfolio, clustering approaches can be useful. Clustering methods divide the whole data points into different clusters based on some similarity measures (e.g., Euclidean distances). *K*-means, hierarchical, self-organizing map (SOM), and fuzzy *C*-means are some of the popular clustering methods which are commonly used in different academic works (Nanda et al. 2010). Liao (2005) carried out a survey of time series clustering procedures which can be performed using real data without any manipulation or using the features extracted from the original data. Recently, Aghabozorgi et al. (2015) reviewed time series clustering. They asserted that every clustering algorithm has its pros and cons which depends on the applications. In other words, there is no panacea clustering method to be useful for all cases. As it has been mentioned by authors, the dynamic changes in stock price time series could be properly recognized by applying shape-based time series clustering.

Furthermore, the similarity measure is the main problematic challenge in clustering time series. Although simple Euclidean distance is widely used in the literature (see, for examples, Irani et al. 2016; Shirkhorshidi et al. 2015), recently many researchers (Bonanno et al. 2001; Basalto et al. 2005) used other measures such as correlation coefficient (Tola et al. 2008) for clustering stock market. D'Urso and Maharaj (2009) suggested a fuzzy approach based on the autocorrelation coefficient of time series which can properly capture the dynamic behavior. The dynamic time warping (DTW) similarity measure which was originally introduced by Berndt and Clifford (1994) was firstly used to explore the temporal pattern of speech (Keogh and Ratanamahatana 2005) and later in other fields (Weng and Liu 2006; Capitani and Ciaccia 2007; Mitsa 2010). Fangwen et al. (2010) suggest a new similarity measure entitled to Sax Feature Vector Model (SFVM) which often yields the same clusters as DTW in less complexity. Jeong et al. (2011) proposed a weighted version of DTW (WDTW) to classify time series.

In this study, our focus is to introduce a superior similarity measure. We also investigate the effect of the introduced similarity measures in time series clustering on the optimality of the obtained portfolio. This is a portfolio in which the gap between the expected (according to optimization model) and the realized mean–variance frontiers has remarkably decreased compared to one constructed using traditional Markowitz's model. In this regard, WDTW and Euclidean distance of

autocorrelation coefficient are considered to be the most suitable similarity measures to cluster stocks before feeding them into Markowitz's optimization model. In detail, after clustering all stocks in different clusters, the representative stock of each cluster is constructed from equal-weighted averaging over its members' stocks. Then, the representative stock is fed to Markowitz's mean–variance optimization model. Consequently, to demonstrate the efficiency of our proposed approach, the cluster-based Markowitz' model is compared to the original Markowitz optimization model which is constructed directly based on the returns and the covariance matrix of whole stocks without implementing any clustering approach and previous work by Tola et al. (2008).

Nevertheless, the Markowitz model is not widely used in practice due to the following reasons (Amenc and Le Sourd 2005; Merton 1980): first, the optimal weights are really sensitive to the mean returns; second, the estimation process of average returns would be statistically a challenging and problematic issue. Raw historical data do not generally constitute a good representation of the forthcoming period. Taillard (2004) also illustrates how sensitive the Markowitz model could be to the data, how it can be affected by error, and how this can lead to an incoherent allocation of stocks in the portfolio. He explained that the optimization produces unstable results in the case of poor conditioning of the variance–covariance matrix. We opine that applying clustering techniques to construct representative stocks would reduce the destructive effects of probable errors in estimating means of individual stocks and lead to a more robust and accurate portfolio optimization.

The paper is organized as follows: in Sect. 2, utilized dataset and preprocessing method as an essential step before clustering are explained. Section 3 is assigned to a brief introduction to the standard Markowitz model and clustering technique. The similarity measures are explained in detail. At the end of this section, a schematic view of the proposed cluster-based methodology is depicted. The performance of the proposed cluster-based portfolio optimization approach is investigated through some real case studies in Sect. 4. The last section presents conclusions and suggestions for future studies.

2 Data and preprocessing

In line with previous studies such as (Tola et al. 2008), a set of daily data of 500 stocks frequently traded at New York Stock Exchange (NYSE) during period 1989–1996 is selected. Furthermore, to prove the validity and reliability of our contributions, the results of implementing our proposed model on a dataset of more recent time periods particularly 2005–2010 (which encompasses financial crises) are provided in Sect. 4.1 and 4.2. We assumed that short selling is not allowed and the returns are quite volatile so that they cannot be forecasted without estimation errors.

Data preprocessing as an important step in data mining techniques is often neglected in financial applications, which might dramatically affect the process of stocks clustering. When there are noisy or missing data, the obtained clusters may not suitably reflect reality (Keogh and Kasetty 2003). Au et al. (2010) say:

Enlightening irregularities is a critical step for data preprocessing.

Before doing the data processing, the daily returns of stocks should be obtained based on their daily adjusted close prices as follows (Amenc and Le Sourd 2005);

$$R_{it} = \log\left(\frac{P_{it}}{P_{i(t-1)}}\right), \tag{1}$$

where R_{it} and P_{it} represent return and adjusted close price of stock i in time (day) t, respectively.

Then, according to the relevant previous studies (Han et al. 2006; Soon and Lee 2007; Witten and Frank 2005), three transformations are utilized to normalize the raw data: "offset translation," "amplitude scaling" and "removing the linear trend."

3 Markowitz's model and clustering technique

3.1 Markowitz's mean–variance model

The main goal of the original Markowitz model is to minimize the portfolio risk for a given value of portfolio expected return. The original Markowitz model (1952, 1959) can be written as:

$$\min \sigma_{\mathrm{p}}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_{ij} \tag{2}$$

$$\text{Subject to;} \quad \sum_{i=1}^{n} w_i R_i \geq R_{\mathrm{p}} \tag{3}$$

$$\sum_{i=1}^{n} w_i = 1 \quad \text{and} \quad w_i \geq 0 \quad \forall i \tag{4}$$

where w_i is the weight of investment in stocks j. R_i is the expected return of the stock i, and R_{p} is the expected return of the portfolio. σ_{ij} is the covariance between stocks i and j; n is the number of stocks.

3.2 Clustering method

Clustering is a class of popular data mining techniques in which similar data elements are placed into groups without having advance knowledge about group allocation. In this study, the average linkage hierarchical clustering (Kaufman and Rousseeuw 2009) as a commonly and powerful technique used in recent years is selected to do time series clustering (Bouguettaya et al. 2015; Keogh and Pazzani 1998; Madhulatha 2012; Murtagh and Contreras 2012; Yim and Ramdeen 2015). In this

regard, a hierarchy of clusters is constructed such that the respective data are not usually partitioned within a single phase. There are two different hierarchical clustering approaches: divisive and agglomerative. The first approach starts with single cluster containing all data points and ends up with a predefined number of clusters. In the second approach, each data point is firstly considered as a single cluster and all these clusters should be properly aggregated to each other in the next step to make new clusters. This process continues until the favorite number of clusters is reached. The detailed algorithm is as below:

Hierarchical Clustering Algorithm

Inputs: $\{X_1,X_2,\ldots,X_s\}$ is a set $\boldsymbol{X}$ of time series, $\boldsymbol{dist(c_i,c_j)}$ is the dissimilarity (for example: the Euclidean distance) between clusters $\boldsymbol{i, j}$ (i.e. the square root of the sum of the square differences between the items they contain) and $\boldsymbol{n}$ is the predefined number of clusters.

Outputs: best agglomerative clusters (C).

```
for i=1 to s
    c_i={X_i}
end for
C={c_1,c_2,…,c_s}
l=s+1
While C.size>n do
(C_min1,C_min2)=minimum dist(c_i,c_j) for all c_i,c_j in C
remove c_min1 and c_min2 from C
add {c_min1,c_min2} to C
end while
```

3.3 Similarity measures

Three sophisticated similarity measures could be used for constructing the cluster-based Markowitz's models which are explained in Table 1 (Liao 2005; Aghabozorgi et al. 2015). The nature of these similarity measures makes it possible to appropriately take volatility and fluctuations of time series into account. These measures are explained in detail in the following subsections.

Table 1 Features and similarity measures used in clustering stocks

Cluster-based Markowitz models	Features	Similarity measures between clusters (each cluster has a representative stock)
Model 1	Autocorrelation coefficients of each stock (with different lags)	Euclidean distance between autocorrelation coefficient series
Model 2	Time series of each stock	Weighted dynamic time warping
Model 3	Time series of each stock	Pearson's correlation coefficient

3.3.1 Autocorrelation coefficient (ACC)

Autocorrelation is a common concept that is known to any graduate student. It is the correlation between all components of a time series for t and $t-l$. In fact, the time lag l indicates how changes in time $t-l$ can affect the future changes in time t. One of the advantages of this approach is the ability of comparing time series with different lengths. (Aghabozorgi et al. 2015; Caiado et al. 2006, 2009).

A set of time series in each cluster which is supposed to follow a similar stochastic process can be mathematically demonstrated as:

$$X = \{x_{st} : s = 1, \ldots, S; t = 1, \ldots, T\} = \begin{pmatrix} x_{11} & \cdots & x_{s1} & \cdots & x_{S1} \\ \vdots & & \vdots & & \vdots \\ x_{1t} & \cdots & x_{st} & \cdots & x_{St} \\ \vdots & & \vdots & & \vdots \\ x_{1T} & \cdots & x_{sT} & \cdots & x_{ST} \end{pmatrix} \tag{5}$$

where x_{st} represents tth occurrence in time series s.

Autocorrelation at lag l ($l=1, \ldots, L=T-1$) is computed as:

$$\rho_{sl} = \frac{\sum_{t=l+1}^{T} \left(x_{st} - \bar{x}_s\right)\left(x_{s(t-l)} - \bar{x}_s\right)}{\sum_{t=1}^{T} \left(x_{st} - \bar{x}_s\right)^2} \tag{6}$$

where $\bar{x}_s$ is the mean of the sth time series.

With regard to the fact that autocorrelation values lie in $[-1, 1]$, Eq. 7 is utilized to scale coefficient as a nonnegative similarity measure.

$$\hat{\rho}_{sl} = \sqrt{2\left(1 - \rho_{sl}\right)}. \tag{7}$$

The feature vector of sth time series with different lags $l=1,\ldots, L=T-1$ is defined as $\hat{\rho}_s = \left(\hat{\rho}_{s1}, \ldots, \hat{\rho}_{sr}, \ldots, \hat{\rho}_{sL}\right)$. This means that the number of features for each time series varies from 1 to $T-1$. Therefore, using these features, the distance between two time series s and r ($d(s,r)$) is calculated as follows:

$$d(s,r) = \sqrt{(\hat{\rho}_s - \hat{\rho}_r)(\hat{\rho}_s - \hat{\rho}_r)'}. \tag{8}$$

3.3.2 Weighted dynamic time warping (WDTW)

In the similarity measure based on dynamic time warping (DTW), every point in a time series could be related to multiple points in another time series and vice versa. In fact, the main advantage of DTW is its capability to align (backward or forward) one point in a time series to multiple points in another one (Berndt and Clifford 1994; Xi et al. 2006). The DTW distance is categorized in shape-based similarity measures that can recognize the temporal shift in an innovative way (Wöllmer et al. 2009). Such as autocorrelation measure, DTW can deal with time

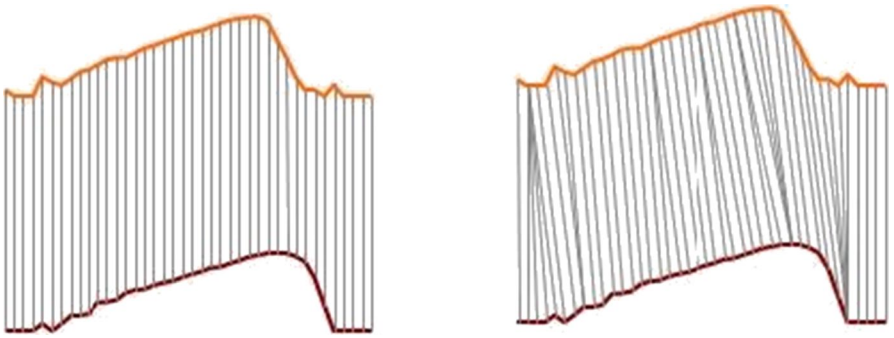

Fig. 1 The left shape is fixed time axis (sequences are aligned "one to one"), and the right shape is warped time axis (nonlinear alignment is possible)

series with different lengths. Figure 1 illustrates the difference between fixed time axis methods such as Euclidean distance and the warped time axis methods such as DTW.

Recently, the weighted version of DTW (WDTW) has been developed for time series clustering to cope with some drawbacks of DTW such as the cases that mapping of one point is performed with more limited points in another time series (Jeong et al. 2011). Suppose $R=\{r_1, r_2,\ldots, r_m\}$ and $S=\{s_1, s_2,\ldots, s_n\}$ are two stocks' return time series with different lengths. The pairwise Euclidean distance of all points of two time series can be calculated as $d(i, j)=(r_i-s_j)^2$ which can be collected in n by m matrix. A feasible alignment between points of two time series called warping

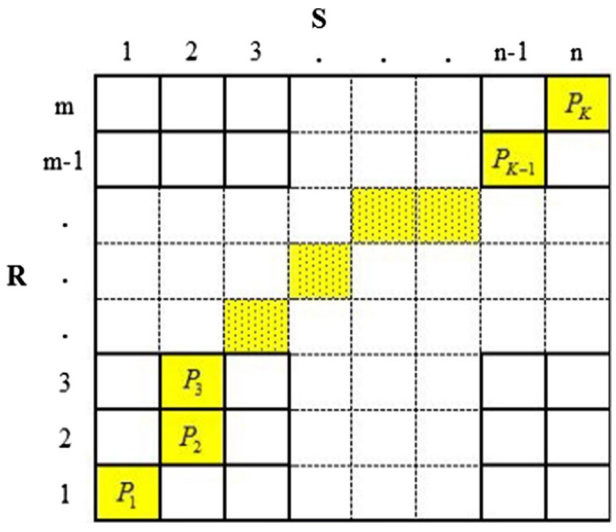

Fig. 2 Color cells specify a hypothetical warping path in m by n pairwise matrix

path (Fig. 2) and is indicated by $P=\{P_1, P_2, \ldots, P_k, \ldots, P_K\}$. It should typically satisfy the following conditions:

Boundary conditions the warping path includes both starting and ending cells in the corresponding pairwise matrix which are indicated as $P_1=[1, 1]$ and $P_K=[m, n]$.
Continuity cells of a path should be adjacent, meaning that if $P_{k-1}=[i_{k-1}, j_{k-1}]$ and $P_k=[i_k, j_k]$, then $i_k - i_{k-1} \leq 1$ and $j_k - j_{k-1} \leq 1$ (i.e., there should not be any discontinuity in the cells of the respective path).
Monotonicity the cells of the respective path should be consecutively progressed, meaning that in case $P_{kv} = [i_{k-1}, j_{k-1}]$ and $P_k = [i_k, j_k]$, then $i_k \geq i_{k-1}$ and $j_k \geq j_{k-1}$ provided that one of these inequalities be a strict inequality.

Any warping path P contains a feasible alignment between two time series R and S with an accumulative distance between aligned points of two time series. The main goal in WDTW is to find the minimum path between two series (optimal warping path, P_{opt}). K in Fig. 2 can be interpreted as the number of features in each comparison of two time series distance.

The WDTW distance can be calculated by the following recursive function:

$$\text{WDTW}(R, S) = \sqrt{\delta(m, n)}, \tag{9}$$

$$\text{where } \delta(i,j) = w(i,j) * d\left(r_i, s_j\right) + \min\{\delta(i-1, j-1),\ \delta(i-1, j),\ \delta(i, j-1)\} \tag{10}$$

$$\text{and } w(i,j) = \left(\frac{w_{\max}}{1 + \exp\left(-g\left(|i-j| - m_c\right)\right)} \right) \tag{11}$$

where $w_{\max}$ is an upper bound for the weight parameter and m_{c} is obtained as follows:

$$m_{\text{c}} = {}^{L}/_{2}, \tag{12}$$

where L is the length of larger time series.

Moreover, g is a constant parameter ranging from zero to infinity that controls the curvature (slope) in Eq. 11 and is used for penalizing the phase (time) difference of points i, j from two time series. As opposed to previous works (such as Jeong et al. 2011; Jeong and Jayaraman 2015) which have considered g as a constant value, in our work, optimal amount of g is obtained using trial and error approach as follows:

$$g = 4.08 * L^{-0.74} - 0.025\ . \tag{13}$$

3.3.3 Pearson's correlation coefficient (PCC)

The correlation coefficient between two time series R and S (which is considered as the clustering similarity measure in this method) is defined as:

$$\rho_{RS} = \frac{\mathrm{Cov}(R, S)}{\sqrt{\mathrm{Var}(R)\mathrm{Var}(S)}} \tag{14}$$

Here, we presume that "time series" are univariate; otherwise the definition does not work as the covariance structure would be a matrix. This measure indicates the linear dependence, and it ranges in $[-1, 1]$. For two independent time series, $\rho = 0$. Equation (7) is also used for scaling correlation coefficient.

3.4 A schematic view of the cluster-based Markowitz models

As mentioned before, after implementing the clustering technique on time series of stocks, a representative stock is considered in place of each cluster (the representative stock of each cluster is constructed from equal-weighted averaging over time series of its members' stocks). At this stage, the Markowitz model can be constructed by feeding the prepared input data including the average returns and covariance matrix of the representative stocks which is expected to show better performance than the feeding individual stocks. Briefly, according to the arguments given above and Fig. 3, we first find representative stocks by clustering stock time series and then feed them to original Markowitz's model to obtain the optimal investment weight of each one.

4 Performance evaluation using experimental results

In this section, the performance of the clustering algorithm used to construct the Markowitz model as well as how different similarity measures can improve the optimization results in term of their accuracy in reflecting the reality (future) is investigated.

The experiments are performed with different number of stocks (50, 100, 150) and time intervals (6 months, 1 and 2 years) using three above-mentioned similarity measures for clustering algorithm. They are performed on two sequences of data: (1) given daily returns for *estimation period* by which the Markowitz optimization model is

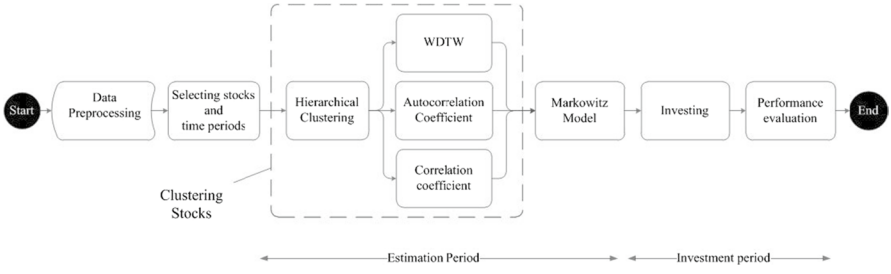

Fig. 3 A brief schematic description of this study

constructed and (2) given daily returns for *investment period* by which the investment is performed and evaluated according to the optimal weights found through the Markowitz optimization model. In other words, the optimal weights of each representative stock (w_i^*) are obtained by solving the optimization model established with the returns vector (R_i) and variance–covariance matrix (σ_{ij}) calculated based on estimation period data.

Given the respective optimal weights, the realized portfolio's return (R_p) and variance (σ_p^2) can be mathematically calculated as follows using the data of the investment period:

$$R_\mathrm{p} = \sum_{i=1}^{N} w_i^* R_i \tag{15}$$

$$\sigma_\mathrm{p}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i^* w_j^* \sigma_{ij}. \tag{16}$$

In the first set of experiments called semi-realistic conditions, we assumed that the investor can forecast R_i accurately. It means that realized returns of stocks are the same as their forecasted returns. In the second set of experiments called full-realistic conditions, this assumption is relaxed.

To establish a reliable measure for comparing the performance of the cluster-based optimization models with that of the original Markowitz's model, the efficient frontier is split to k equal-distance points starting from global minimum variance portfolio (GMVP) and ending to the point with maximum return. Hence, the respective measure can be defined as follows:

$$\tau = \frac{\sum_{k=1}^{K} \sqrt{\left|\frac{R_\mathrm{p}(k)-\hat{R}_\mathrm{p}(k)}{\hat{R}_\mathrm{p}(k)}\right|^2 + \left|\frac{\sigma_\mathrm{p}(k)-\hat{\sigma}_\mathrm{p}(k)}{\hat{\sigma}_\mathrm{p}(k)}\right|^2}}{K} \tag{17}$$

$R_\mathrm{p}(k)$ and $\sigma_\mathrm{p}(k)$ are the return and the standard deviation of point k on the realized efficient frontier obtained based on investment period data. Similarly, $\hat{R}_\mathrm{p}(k)$ and $\hat{\sigma}_\mathrm{p}(k)$ are the return and standard deviation of point k on efficient frontier obtained from the data of the estimation period. In fact, τ measures the gap between what an investor expects and what he/she observes in reality.

Here, the whole efficient frontier is divided into ten equally spaced parts (i.e., $K=10$). It is also obvious that a smaller τ implies a more reliable model in which the reality can be more accurately represented via the respective mathematical model.

In order to show the robustness and accuracy of our proposed approach, each experiment is repeated twenty times for each predefined number of randomly selected stocks (50, 100, 150) using the data of estimation and investment period.

4.1 Cluster-based portfolio optimization under semi-realistic conditions

Under semi-realistic condition, forecasted returns are assumed to be the same as realized returns $\left(R_{\mathrm{p}} = \hat{R}_{\mathrm{p}}\right)$, so Eq. 17 is simplified as follows:

$$\tau = \frac{\sum_{k=1}^{K} \left| \frac{\sigma_{\mathrm{p}}(k) - \hat{\sigma}_{\mathrm{p}}(k)}{\hat{\sigma}_{\mathrm{p}}(k)} \right|}{K}. \tag{18}$$

It says that only the forecasted and realized variances of the portfolios need to be compared. This case of mean–variance model is called minimum variance portfolio (MVP) in the literature.

Tables 2, 3 and 4 summarize obtained results from using varied number of stocks (50, 100 and 150) and time periods [125 days (6 months), 250 days (1 year) and 500 days (2 years)]. For instance, in the case of time period 250, the year of 1989 is considered as estimation period to calculate forecasted portfolio's risk, and the year of 1990 is considered as investment period to compute realized portfolio's risk. It is worth mentioning that the superior result among four presented models are represented in a bold color for all different experiments.

As it is obvious in Tables 2, 3 and 4, cluster-based portfolio optimization models often (in most cases) outperform original Markowitz's minimum variance portfolio model. There are only three cases (out of all 81 experiments in 1989–1996) and zero cases (out of all 45 experiments in 2005–2010) that the original Markowitz model shows better results than others. Furthermore, ACC-based model ends up with better results than WDTW- and WDTW-based model mostly lead to superior results compared to correlation coefficient one. One reason that the two first similarity measures usually lead to superior results than the correlation coefficient is that in high volatility periods the cross-correlation is very unstable measure of similarity. As illustrated in the results of experiments, it is very rare that the using correlation coefficient for clustering leads to proper results in comparison with ACC and WDTW especially in financial crises period. In general, ACC-based model produces remarkably more accurate results than other methods as the number of stocks increases.

As a case in point, to investigate the efficiency of these cluster-based models, 100 stocks are randomly selected from 500 stocks. The time periods 1991 and 1992 are used as estimation and investment period, respectively. According to Fig. 4, there is a remarkably smaller gap between the frontiers obtained from estimation period and investment period for cluster-based optimization models. This is also more obvious for ACC-based model that the gap is negligible. However, as it is expected and has been most frequently referred in the literature, the respective gap is considerably high for original Markowitz's model.

4.2 Cluster-based portfolio optimization under full-realistic conditions

Under full-realistic conditions, it is assumed that both future returns and risk should be forecasted (there is no previous knowledge about the future). It means that both

Table 2 Reliability measure τ of Eq. (18) for 6 months' time period under semi-realistic conditions

Est. period	Inv. period	50 stocks				100 stocks				150 stocks			
		MAR	Clustering			MAR	Clustering			MAR	Clustering		
			ACC	WDTW	PCC		ACC	WDTW	PCC		ACC	WDTW	PCC
89(1)	89(2)	0.420	**0.260**	0.305	0.323	0.530	0.372	0.423	**0.354**	0.613	**0.456**	0.527	0.475
89(2)	90(1)	0.223	**0.132**	0.146	0.145	0.267	**0.164**	0.265	0.183	0.281	**0.169**	0.272	0.170
90(1)	90(2)	0.577	**0.517**	0.628	0.528	0.878	0.700	0.716	**0.647**	1.022	0.701	**0.667**	0.673
90(2)	91(1)	**0.107**	0.109	0.134	0.192	0.112	**0.099**	0.115	0.104	0.116	0.123	0.126	**0.114**
91(1)	91(2)	0.156	0.126	**0.109**	0.134	0.221	**0.125**	0.149	0.129	0.222	0.132	0.138	**0.113**
91(2)	92(1)	0.242	0.201	**0.156**	0.221	0.359	**0.157**	0.231	0.199	0.392	**0.200**	0.236	0.217
92(1)	92(2)	0.227	0.246	**0.159**	0.186	0.276	**0.125**	0.197	0.208	0.273	**0.152**	0.243	0.169
92(2)	93(1)	0.389	**0.217**	0.329	0.257	0.644	**0.269**	0.348	0.274	0.781	**0.287**	0.402	0.431
93(1)	93(2)	0.159	**0.070**	0.093	0.074	0.286	0.183	0.176	**0.126**	0.294	**0.095**	0.142	0.107
93(2)	94(1)	0.414	0.574	0.401	**0.361**	0.623	0.511	**0.468**	0.488	0.749	**0.517**	0.532	0.547
94(1)	94(2)	0.216	0.121	0.190	**0.084**	0.255	**0.103**	0.257	**0.103**	0.376	**0.112**	0.394	0.146
94(2)	95(1)	0.137	0.114	**0.111**	0.111	0.253	0.122	0.121	**0.077**	0.279	0.169	0.149	**0.100**
95(1)	95(2)	0.263	0.168	**0.164**	0.170	0.362	0.209	**0.184**	0.225	0.499	0.287	0.252	**0.197**
95(2)	96(1)	0.508	0.438	**0.399**	0.421	0.705	**0.536**	0.614	0.603	0.766	**0.522**	0.628	0.598
96(1)	96(2)	0.167	**0.088**	0.141	0.138	0.277	0.169	**0.101**	0.145	0.331	0.189	0.199	**0.171**
2005(1)	2005(2)	0.475	0.330	**0.279**	0.289	0.169	**0.144**	0.180	0.153	0.237	**0.152**	0.175	0.181
2005(2)	2006(1)	0.473	0.326	**0.278**	0.282	0.553	0.339	0.436	**0.290**	0.322	0.246	**0.169**	0.216
2006(1)	2006(2)	0.116	**0.109**	0.124	0.113	0.116	0.107	**0.098**	0.107	0.094	0.082	**0.078**	0.081
2006(2)	2007(1)	0.322	**0.251**	0.271	0.268	0.610	**0.397**	0.552	0.559	0.637	**0.456**	0.536	0.532
2007(1)	2007(2)	0.519	0.559	**0.505**	0.512	0.875	0.840	**0.759**	0.782	0.986	**0.800**	0.827	0.815
2007(2)	2008(1)	0.129	0.094	**0.065**	0.076	0.188	**0.147**	0.182	0.155	0.249	**0.173**	0.175	0.177
2008(1)	2008(2)	2.084	**1.701**	1.899	1.829	2.142	**1.768**	1.898	1.931	2.390	**2.016**	2.161	2.075
2008(2)	2009(1)	0.236	0.264	0.247	**0.229**	0.335	**0.319**	0.363	0.328	0.204	0.210	**0.191**	0.193

Table 2 (continued)

Est. period	Inv. period	50 stocks				100 stocks				150 stocks			
		MAR	Clustering			MAR	Clustering			MAR	Clustering		
			ACC	WDTW	PCC		ACC	WDTW	PCC		ACC	WDTW	PCC
2009(1)	2009(2)	0.362	0.348	**0.339**	0.341	0.522	**0.425**	0.505	0.425	0.429	**0.399**	0.443	0.403

89(1): first 6 month of 1989 and 89(2): second 6 month of 1989 under semi-realistic conditions (MAR = original Markowitz, auto-ACC = autocorrelation coefficient, WDTW = weighted dynamic time warping and PCC = Pearson's correlation coefficient)

Table 3 Reliability measure τ of Eq. (18) for 1-year time period under semi-realistic conditions

Est. period	Inv. period	50 stocks				100 stocks				150 stocks			
		MAR	Clustering			MAR	Clustering			MAR	Clustering		
			ACC	WDTW	PCC		ACC	WDTW	PCC		ACC	WDTW	PCC
1989	1990	0.410	0.398	**0.375**	0.384	0.428	0.402	**0.378**	0.385	0.487	0.432	**0.407**	0.426
1990	1991	0.115	0.108	0.133	**0.103**	0.119	**0.100**	0.105	0.110	0.112	0.201	**0.088**	0.100
1991	1992	0.231	**0.098**	0.116	0.107	0.214	**0.095**	0.120	0.118	0.222	**0.090**	0.156	0.131
1992	1993	0.170	**0.105**	0.160	0.115	0.259	0.156	0.171	**0.107**	0.214	0.182	0.148	**0.114**
1993	1994	0.359	**0.273**	0.308	0.317	0.357	**0.284**	0.302	0.301	0.401	**0.263**	0.437	0.313
1994	1995	0.159	0.153	0.148	**0.116**	0.155	**0.098**	0.135	0.155	0.261	**0.129**	0.204	0.161
1995	1996	0.429	0.334	**0.318**	0.387	0.475	0.424	**0.378**	0.436	0.540	**0.439**	0.448	0.449
2005	2006	0.175	**0.076**	0.148	0.096	0.175	0.148	**0.137**	0.149	0.137	0.163	**0.132**	0.205
2006	2007	0.510	**0.483**	0.508	0.617	0.648	**0.553**	0.636	0.572	0.688	**0.559**	0.639	0.624
2007	2008	1.716	**1.441**	1.663	1.482	1.771	1.566	**1.561**	1.570	1.836	**1.610**	1.688	1.789
2008	2009	0.315	**0.266**	0.280	0.313	0.299	0.324	**0.278**	0.321	0.284	0.317	**0.264**	0.269

Table 4 Reliability measure τ of Eq. (18) for 2 years' time period under semi-realistic conditions

Est. period	Inv. period	50 stocks				100 stocks				150 stocks			
		MAR	Clustering			MAR	Clustering			MAR	Clustering		
			ACC	WDTW	PCC		ACC	WDTW	PCC		ACC	WDTW	PCC
89–90	91–92	0.182	**0.115**	0.150	0.154	0.192	**0.109**	0.206	0.200	0.272	**0.130**	0.220	0.224
90–91	92–93	0.125	**0.117**	0.138	0.160	**0.110**	0.157	0.140	0.163	**0.109**	0.147	0.151	0.187
91–92	93–94	0.219	0.157	**0.135**	0.137	0.229	**0.121**	0.151	0.128	0.267	**0.110**	0.136	0.137
92–93	94–95	0.164	0.096	0.127	**0.075**	0.173	**0.109**	0.114	0.114	0.215	0.172	0.190	**0.107**
93–94	95–96	0.085	0.099	**0.076**	0.078	0.156	0.095	0.131	**0.094**	0.242	**0.081**	0.254	0.084
2005–2006	2007–2008	1.824	**1.549**	1.716	1.730	1.810	1.594	**1.535**	1.674	2.209	**1.820**	1.962	1.867
2007–2007	2008–2009	1.622	**1.265**	1.273	1.432	1.659	1.374	**1.357**	1.417	1.673	1.388	**1.327**	1.487

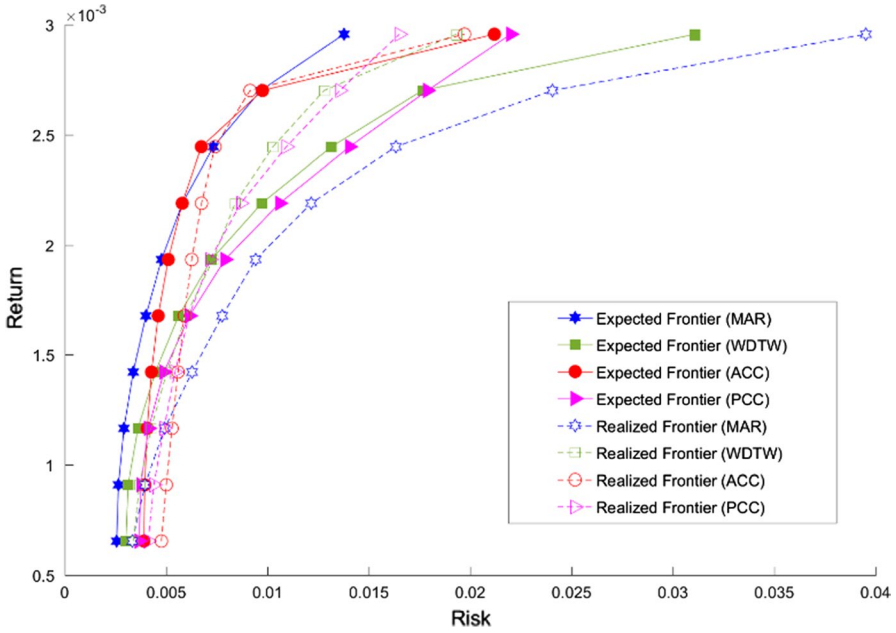

Fig. 4 The filled marker with solid lines is the expected frontiers, and the unfilled marker with dashed lines is the realized frontiers in the reality (investment period). The blue (star marker), green (square marker), red (circle marker) and pink (triangle marker) colors refer to the Markowitz model, WDTW, ACC and PCC based clustering methods, respectively. (Color figure online)

Table 5 Reliability measure τ of Eq. (17) for 1-year time period under full-realistic conditions

Est. period	Inv. period	100 stocks				150 stocks			
		MAR	Clustering			MAR	Clustering		
			ACC	WDTW	PCC		ACC	WDTW	PCC
1989	1990	1.188	1.188	**1.175**	1.215	1.237	**1.107**	1.234	1.258
1990	1991	**0.484**	0.697	0.996	0.961	**0.373**	0.762	0.798	0.660
1991	1992	0.690	0.593	**0.540**	0.604	0.539	**0.429**	0.491	0.445
1992	1993	0.545	0.449	**0.415**	0.443	0.556	0.418	**0.398**	0.405
1993	1994	1.102	1.137	**1.101**	1.193	0.840	0.839	**0.831**	0.834
1994	1995	**0.721**	1.346	1.234	1.424	0.011	0.116	**0.008**	0.012
1995	1996	0.773	0.611	0.625	**0.577**	0.856	0.683	**0.659**	0.717
2005	2006	0.267	0.179	**0.174**	0.198	0.231	0.142	**0.118**	0.160
2006	2007	0.778	0.727	**0.673**	0.693	1.527	1.549	**1.448**	1.798
2007	2008	0.930	1.019	**0.929**	1.020	0.996	0.966	**0.902**	1.086
2008	2009	0.218	0.326	**0.179**	0.296	3.740	**2.681**	7.452	6.593

Table 6 Reliability measure τ of Eq. (17) for 2-year time periods under full realistic conditions

Est. Period	Inv. period	100 stocks				150 stocks			
		MAR	Clustering			MAR	Clustering		
			ACC	WDTW	PCC		ACC	WDTW	PCC
89–90	91–92	0.309	0.284	**0.271**	0.293	0.331	0.318	**0.303**	0.320
90–91	92–93	0.510	0.402	**0.386**	0.520	0.529	0.495	**0.402**	0.458
91–92	93–94	0.872	0.796	**0.794**	0.804	0.880	0.829	**0.807**	0.812
92–93	94–95	0.712	0.627	**0.575**	0.621	0.727	0.574	**0.539**	0.646
93–94	95–96	0.693	0.566	**0.550**	0.584	0.667	0.503	0.534	**0.489**
2005–2006	2007–2008	2.835	**2.731**	2.906	2.857	2.978	**2.841**	3.116	2.844
2007–2007	2008–2009	1.512	1.574	**1.503**	1.527	1.459	1.410	**1.409**	1.447

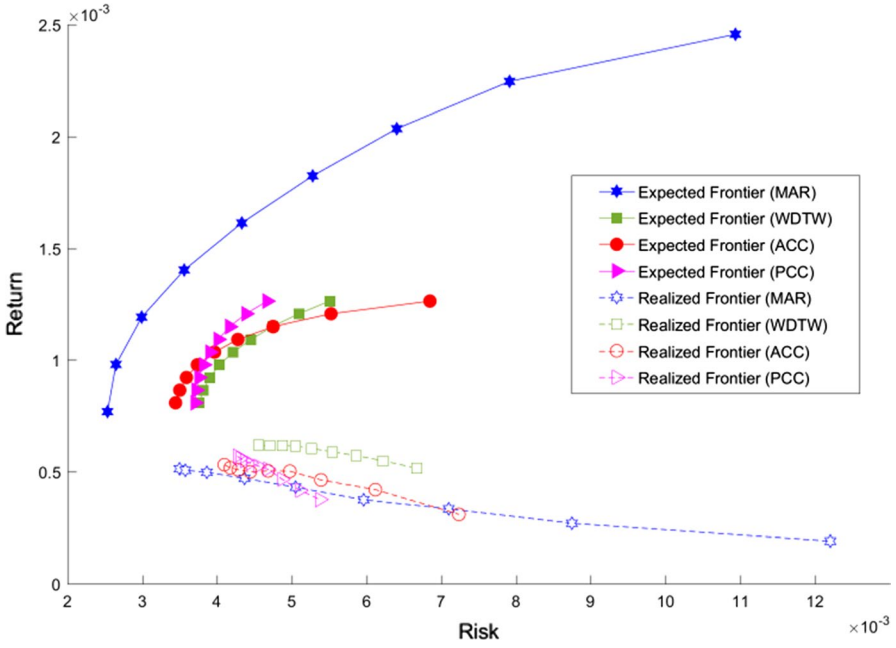

Fig. 5 The filled marker with solid lines is the expected frontiers, and the unfilled marker with dashed lines is the realized frontiers in the reality (investment period). The blue (star marker), green (square marker), red (circle marker) and pink (triangle marker) colors refer to the Markowitz model, WDTW, ACC and PCC based clustering methods, respectively. (Color figure online)

these parameters obtained through the optimization and the simulation routines should be taken into account to assess the performance of cluster-based models. The numbers of stocks in these experiments are 100 and 150 under two different time periods: 250 and 500.

Tables 5 and 6 demonstrated that the WDTW-based model is mostly superior to others. The statistics show the dominance of cluster-based portfolio optimization (especially the WDTW-based model) over the original Markowitz model because there are only three cases (out of all 29 experiments in 1989–1996) and zero cases (out of all 12 experiments in 2005–2010) that original Markowitz model shows better results than others.

This can also be inferred from Fig. 5 where the experiments are performed for 150 stocks in time period 1992–1995. The gap for all cluster-based models is remarkably low compared to original Markowitz model. The interesting point is that the results obtained using cluster-based portfolio optimization models can outperform the results achieved based on the original Markowitz model for all different implemented experiments (Table 6).

5 Conclusion

To obtain a more robust and accurate portfolio through the original Markowitz's optimization model, we apply two time series clustering similarity measures: autocorrelation coefficients (ACC) and weighted dynamic time warping (WDTW) in addition to the popular measure of Pearson's correlation coefficient (PCC). The hierarchical clustering as a well-known technique was used for clustering time series of stocks. We assumed that the return series of each cluster (i.e., representative stock) can be constructed using equally weighted average of all returns of stocks belonging to the respective cluster. To show the effectiveness of our proposed model and compare our results with those of previous researches, 500 stocks were selected which had been frequently traded at New York Stock Exchange (NYSE) for two time periods of 1989–1996 and 2005–2010. All experiments are performed in two different market conditions: semi-realistic and full-realistic. In the first one, the investor is assumed to have perfect information about the future return of stocks, that is, the investor knows the returns but should forecast the covariance matrix. In the second one, both futures return and the covariance matrix of stocks need to be estimated. The experimental results in both states verified that the new clustering-based portfolio optimization techniques could achieve a remarkable improvement compared to both previous cluster-based methods using Pearson's correlation coefficient as a similarity measure and the original Markowitz model in terms of the gap between what an investor expects and what he/she observes in reality.

Acknowledgements We are grateful to Institute for Plasma Research of Kharazmi University for all their kindness and help in terms of providing us with their super computer and facilities.

References

Aghabozorgi S, Shirkhorshidi AS, Wah TY (2015) Time-series clustering—a decade review. Inf Syst 53:16–38

Amenc N, Le Sourd V (2005) Portfolio theory and performance analysis. Wiley, Hoboken

Au S-T, Duan R, Hesar SG, Jiang W (2010) A framework of irregularity enlightenment for data preprocessing in data mining. Ann Oper Res 174:47–66
Basalto N, Bellotti R, De Carlo F, Facchi P, Pascazio S (2005) Clustering stock market companies via chaotic map synchronization. Phys A 345:196–206
Berndt DJ, Clifford J (1994) Using dynamic time warping to find patterns in time series. In: KDD workshop, vol 16. Seattle, WA, pp 359–370
Best MJ (2010) Portfolio optimization. CRC Press, Boca Raton
Bonanno G, Lillo F, Mantegna RN (2001) High-frequency cross-correlation in a set of stocks. Quant Finance 96:104
Bouguettaya A, Yu Q, Liu X, Zhou X, Song A (2015) Efficient agglomerative hierarchical clustering. Expert Syst Appl 42:2785–2797
Caiado J, Crato N, Peña D (2006) A periodogram-based metric for time series classification. Comput Stat Data Anal 50:2668–2684
Caiado J, Crato N, Peña D (2009) Comparison of times series with unequal length in the frequency domain. Commun Stat Simul Comput 38:527–540
Capitani P, Ciaccia P (2007) Warping the time on data streams. Data Knowl Eng 62:438–458
Cong F, Oosterlee C (2016) Multi-period mean-variance portfolio optimization based on monte-carlo simulation. J Econ Dyn Control 64:23–38
D'Urso P, Maharaj EA (2009) Autocorrelation-based fuzzy clustering of time series. Fuzzy Sets Syst 160:3565–3589
Ehrgott M, Klamroth K, Schwehm C (2004) An MCDM approach to portfolio optimization. Eur J Oper Res 155:752–770
Fangwen Z, Zehong Y, Yixu S, Yi L (2010) A novel similarity measure framework on financial data mining. In: 2010 second international conference on networks security wireless communications and trusted computing (NSWCTC). IEEE, pp 505–508
Han J, Kamber M, Pei J (2006) Data mining, southeast Asia edition: concepts and techniques. Morgan Kaufmann, Burlington
Irani J, Pise N, Phatak M (2016) Clustering techniques and the similarity measures used in clustering: a survey. Int J Comput Appl 134:9–14
Jeong Y-S, Jayaraman R (2015) Support vector-based algorithms with weighted dynamic time warping kernel function for time series classification. Knowl Based Syst 75:184–191
Jeong Y-S, Jeong MK, Omitaomu OA (2011) Weighted dynamic time warping for time series classification. Pattern Recognit 44:2231–2240
Kaufman L, Rousseeuw PJ (2009) Finding groups in data: an introduction to cluster analysis, vol 344. Wiley, Hoboken
Keogh E, Kasetty S (2003) On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Min Knowl Disc 7:349–371
Keogh EJ, Pazzani MJ (1998) An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In: KDD, pp 239–243
Keogh E, Ratanamahatana CA (2005) Exact indexing of dynamic time warping. Knowl Inf Syst 7:358–386
Liao TW (2005) Clustering of time series data—a survey. Pattern Recognit 38:1857–1874
Lin C-C, Liu Y-T (2008) Genetic algorithms for portfolio selection problems with minimum transaction lots. Eur J Oper Res 185:393–404
Madhulatha TS (2012) An overview on clustering methods. arXiv preprint arXiv:12051117
Markowitz H (1952) Portfolio selection. J Finance 7:77–91
Markowitz HM (1959) Portfolio selection: efficient diversification of investments, 2nd edn. Wiley, Hoboken
Merton RC (1980) On estimating the expected return on the market: an exploratory investigation. J Financ Econ 8:323–361
Mitsa T (2010) Temporal data mining. CRC Press, Boca Raton
Murtagh F, Contreras P (2012) Algorithms for hierarchical clustering: an overview. Wiley Interdiscip Rev Data Min Knowl Discov 2:86–97
Nanda S, Mahanty B, Tiwari M (2010) Clustering Indian stock market data for portfolio management. Expert Syst Appl 37:8793–8798
Sharpe WF (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Finance 19:425–442

Shirkhorshidi AS, Aghabozorgi S, Wah TY (2015) A comparison study on similarity and dissimilarity measures in clustering continuous data. PLoS ONE 10:e0144059

Soon L-K, Lee SH (2007) An empirical study of similarity search in stock data. In: Proceedings of the 2nd international workshop on Integrating artificial intelligence and data mining, vol 84. Australian Computer Society, Inc., pp 31–38

Taillard G (2004) Le point sur? L'optimisation de portefeuille. Bankers, Markets & Investors No. 65

Tola V, Lillo F, Gallegati M, Mantegna RN (2008) Cluster analysis for portfolio optimization. J Econ Dyn Control 32:235–258

Vaclavik M, Jablonsky J (2012) Revisions of modern portfolio theory optimization model. CEJOR 20:473–483

Weng S-S, Liu Y-H (2006) Mining time series data for segmentation by using ant colony optimization. Eur J Oper Res 173:921–937

Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington

Wöllmer M, Al-Hames M, Eyben F, Schuller B, Rigoll G (2009) A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams. Neurocomputing 73:366–380

Xi X, Keogh E, Shelton C, Wei L, Ratanamahatana CA (2006) Fast time series classification using numerosity reduction. In: Proceedings of the 23rd international conference on machine learning. ACM, pp 1033–1040

Yim O, Ramdeen KT (2015) Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. Quant Methods Psychol 11:8–21