Soft Computing
https://doi.org/10.1007/s00500-025-10656-2

FOCUS

Time series clustering for high-dimensional portfolio selection: a comparative study

Raffaele Mattera[1] · Germana Scepi[2] · Parmjit Kaur[3,4]

Accepted: 18 January 2025

Abstract
In high-dimensional portfolio selection, traditional asset allocation techniques often yield suboptimal results out-of-sample, while equally weighted portfolios have shown better performances in such scenarios. To leverage the advantages of diversification while addressing the curse of dimensionality, we turn to clustering techniques. Specifically, we explore the application of k-means clustering for time series, which offers a clear financial interpretation as the prototype of each cluster represents an equally weighted portfolio of the assets within the cluster. In this paper, we conduct a comprehensive comparison of various time series clustering techniques in the context of portfolio performance. By evaluating the out-of-sample performance of portfolios constructed using different clustering approaches, we aim to identify the most effective method for investment purposes.

Keywords Cluster analysis · Finance · Financial time series · Unsupervised learning · Asset allocation

1 Introduction

Clustering, a cornerstone algorithm in data mining, finds wide application in exploratory analysis, anomaly detection, and classification tasks. However, its utility extends even further when dealing with time series data, albeit with added complexity owing to the challenge of defining similarity measures.

In the domain of finance, where the clustering of time series such as stock prices and returns is commonplace, consideration of empirical regularities-commonly referred to as *stylized facts*-is needed (Bastos and Caiado 2021). For instance, stock prices typically exhibit integrated behavior, necessitating the use of returns for clustering analyses. Moreover, the presence of volatility clustering, wherein volatility tends to cluster in groups of low/high values over time, poses a further challenge. Additionally, the non-Gaussian, asymmetric, and heavy-tailed nature of returns' empirical densities complicates clustering efforts.

In financial markets, a natural application of time series clustering techniques is found in portfolio construction, also known as asset allocation task (Mantegna 1999; Caiado and Crato 2010; Iorio et al. 2018; Raffinot 2017). Given the prevalence of high-dimensional data, wherein the number of available assets N often surpasses the number of time observations T, traditional optimization strategies face challenges in estimating the inverse of asset returns' covariance matrix, leading to increased estimation errors (e.g. see Ledoit and Wolf 2017). In such context, naive equally weighted portfolios outperform optimal portfolio allocation strategies, such as the mean-variance or the global minimum-variance portfolios (DeMiguel et al. 2009).

In econophysics literature, clustering emerged as a powerful tool in high-dimensional setting, facilitating the identification of smaller sets of stocks to construct diversified funds, and researchers have proposed various methods to address this issue. For instance, Tola et al. (2008) highlighted the significance of clustering time series data in finance for

✉ Raffaele Mattera
raffaele.mattera@unicampania.it

Germana Scepi
scepi@unina.it

Parmjit Kaur
p.kaur@londonmet.ac.uk

[1] Department of Mathematics and Physics, University of Campania"Luigi Vanvitelli", Caserta, Italy

[2] Department of Economics and Statistics, University of Naples "Federico II", Naples, Italy

[3] Guildhall School of Business and Law, London Metropolitan University, London, UK

[4] Babes-Bolyai University, Cluj-Napoca, Romania

Published online: 17 May 2025

identifying similarities among financial assets, crucial for portfolio optimization and risk management. Tumminello et al. (2007) discussed correlation-based networks of equity returns sampled at different time horizons, demonstrating how clustering analysis can derive clusters of stocks from price return time series. Clustering approaches based on correlation matrix have also been exploited with similar purposes in other domains (e.g. see Ausloos and Lambiotte 2007; Gligor and Ausloos 2007). In the financial setting, Raffinot (2017) proposed the use of correlation-based distance in hierarchical clustering algorithm for portfolio selection. More recently, Iorio et al. (2018) proposed to build portfolios of stock with a P-spline-based distance, while Lorenzo and Arroyo (2023) proposed a clustering approach based on risk/return trade-off for building portfolios of cryptocurrencies. Furthermore, innovative optimization approaches exploiting clustering have emerged for portfolio selection. For instance, Soleymani and Vasighi (2022) employed the k-means++ clustering technique to cluster financial returns, subsequently using risk measures like CVaR to identify the riskiest groups of stocks. Additionally, Gubu and Rosadi (2020) compared Kamila and Weighted K-Mean clustering for robust mean-variance portfolio selection, emphasizing the role of cluster analysis in creating efficient portfolios. More recently, Mattera et al. (2024) implemented a forecast-based investment strategy based on time series clustering.

Overall, the application of econophysics in investigating time series clustering for high-dimensional portfolio selection involves a diverse range of methodologies, all aimed at enhancing portfolio optimization strategies. The k-means clustering algorithm is one of the most popular method in this regard, wherein the prototype can be interpreted as the equally weighted portfolio of the assets included in each cluster. Equally weighted portfolios are particularly effective in high-dimensional settings where N is large, as they require no estimation (DeMiguel et al. 2009). However, as highlighted in the previously mentioned papers (e.g.]raffinot2017hierarchical,iorio2018,lorenzo2023online, for a smaller set of K equally weighted funds, optimal diversification strategies such as mean-variance and GMV (Global Minimum Variance) can be employed with success. These strategies offer indeed a more refined approach to portfolio construction, taking into account the covariance structure of assets and optimizing the risk-return trade-off.

Therefore, to leverage the advantages of diversification while addressing the dimensionality challenge, it is convenient to use clustering techniques. However, it is still not clear what approach provides the best performances in terms of out-of-sample return/risk trade-off, as most of previous studies compared the clustering-based portfolio strategies with others not involving clustering. Differently, in this paper we conduct a comparison of various time series clustering techniques in the context of portfolio performance. By evaluating the out-of-sample performance of portfolios constructed using different clustering approaches, we aim to identify the most effective method for investment purposes. Our analysis sheds light on the relationship between clustering techniques and portfolio outcomes, providing valuable insights for investors and portfolio managers.

The empirical analysis focuses on two datasets, representing both high-dimensional and low-dimensional portfolios. Precisely, we build portfolios on Dow Jones Industrial Average constituents and to those included in the Russell 200 Index. The strategies are evaluated under alternative performance metrics. In sum, we find that the use of clustering is particularly beneficial in high-dimensional asset allocation, since clustered portfolios perform much better than standard approaches used by practitioners in high-dimensional settings. However, the use of clustering does not allow for relevant benefits when a small number of assets universe is considered.

The remainder of this paper is organized as follows: Section 2 provides a review of clustering methodologies in finance; Section 3 outlines the methodology employed in our study; Section 4 presents the empirical analysis and results; and Section 5 concludes the paper and suggests avenues for future research.

2 Measuring the proximity of financial time series

Once a dissimilarity measure between the objects has been specified, a clustering algorithm must be chosen to obtain the partitions. As stated by Liao (2005), most clustering techniques for time series *"try to modify the existing algorithms for clustering static data in such a way that time series can be handled"*. This is usually done using proper time series distance matrices. In what follows, we provide a brief overview of the most commonly employed distance measures for financial time series.

Given a pair of stocks returns' time series $r_{n,t}$ and $r_{n\prime,t}$, a first approach for clustering the time series could be simply the Euclidean distance between the two raw time series:

$$d_{\text{EUCL}_{n,n\prime}} = \sqrt{\sum_{t=1}^{T}(r_{n,t} - r_{n\prime,t})^2} \tag{1}$$

There are at different reasons why this measure is inadequate for clustering financial time series. First, (1) does not account for the serial correlation structure of the data and ignores the correlation structure of the assets, a crucial aspect of portfolio selection. Moreover, the simple Euclidean distance on temporal ordinates (1) may be less effective

in clustering illiquid stocks characterized by sparse returns matrix.

Starting from these limitations, Mantegna (1999) and Raffinot (2017) proposed to quantify the dissimilarity among different stocks according to their estimated correlation coefficient. The simple difference between estimated correlation coefficients cannot be used as a distance since it does not fulfil the axioms that define a metric. To overcome this issue, Mantegna (1999) proposed to use the distance

$$d_{\mathrm{COR}_{n,n\prime}} = \sqrt{2(1-\rho_{n,n\prime})}, \tag{2}$$

that depends by the correlation $\rho_{n,j}$ between the n-th stock returns $r_{n,t}$ and the $n\prime$-th returns $r_{n\prime,t}$. However, the correlation coefficient is still a static measure that does not properly account for the dynamic structure of the time series.

Interesting approaches are based on the frequency domain representation of the time series (Díaz and Vilar 2010), which measures the similarity of two n and $n\prime$ time series in terms of their spectral densities. An unbiased estimator of the actual spectral density is the periodogram

$$I_n(\lambda_l) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} r_{n,t}e^{-i\lambda t}\right|^2 \quad \lambda \in [-\pi,\pi],$$

at frequencies $\lambda_l = 2\pi l/T$, given $\{l = -L,\ldots,L\}$ with $L = (T-1)/2$. Given σ_n^2 be the sample variance of $r_{n,t}$, Caiado et al. (2006) proposed to consider the following distance between the log-normalized periodograms

$$d_{\mathrm{NPER}_{n,n\prime}} = \sqrt{\sum_{l=1}^{L}\left(\log\frac{I_n(\lambda_l)}{\sigma_n^2} - \log\frac{I_{n\prime}(\lambda_l)}{\sigma_{n\prime}^2}\right)^2}. \tag{3}$$

This approach has been considered in Caiado and Crato (2010) for clustering financial time series. Cepstral coefficients could also be used (D'Urso et al. 2020), although their use requires additional computational challenges.

Within the class of feature-based approaches for financial time series, long memory-based procedures have been recently proposed (Cerqueti and Mattera 2023; Di Sciorio 2023; Lahmiri 2016; Mahmoudi 2021). Long memory, also known as long-range dependence, signifies the presence of enduring, self-correlated patterns in a time series (Ausloos et al. 2017). It implies that past occurrences significantly influence future trends rendering, in the case of financial market, the dynamics of the system under study more than a mere sequence of random data points (Ausloos 2000). Notably previous research highlight the correlation between long memory and the predictability of assets (Cerqueti and Fanelli 2021; Vogl 2023), while other studies shown that long memory can be used for either building profitable investment strategies (Ausloos and Bronlet 2003; Ramos-Requena et al. 2017), to investigate market efficiency (Dimitrova et al. 2019; Mattera et al. 2022) and to better pricing options (Mattera and Di Sciorio 2021). The Hurst exponent of an asset n, denoted as h_n, serves as a crucial metric for quantifying long memory or long-range dependence within time series data. A higher Hurst exponent, typically exceeding 0.5, indicates persistent behavior, suggesting that the financial time series adheres to long-term trends and displays enduring correlations. Conversely, a Hurst exponent lower than 0.5 suggests anti-persistent behavior, wherein the time series tends to revert to the mean and exhibits short-term correlations (Sánchez et al. 2015). From the clustering perspective, we notice that different approaches for estimating the Hurst exponent are available. We can define $h_{n,p}$ as the Hurst exponent estimated using the p-th available approach. Lahmiri (2016) suggests that the set of different Hurst estimators provides a general characterization of the underlying financial time series generating process. Therefore, a simple Hurst-based clustering approach is to consider the Euclidean distance between n and $n\prime$ based on the different Hurst estimates obtained with alternative estimators, that is

$$d_{\mathrm{HURST}_{n,n\prime}} = \sqrt{\sum_{p=1}^{P}(h_{n,p} - h_{n\prime,p})^2} \tag{4}$$

A widely used alternative to future-based approaches are the so-called model-based approaches. This class of clustering models assume that the returns in time series are generated by a specific statistical model, such that we can measure the proximity between two time series by the similarity of the fitted models. In this respect, an important contribution has been proposed by Piccolo (1990), that defined a metric in the class of invertible ARIMA processes as the Euclidean distance between the AR(∞) representation of a given stock returns series $r_{n,t}$. In practice, we compute an AR(P) representation where P is selected according to information criteria. Then, the resulting AR-distance of Piccolo (1990) is

$$d_{\mathrm{ARMA}_{n,n\prime}} = \sqrt{\sum_{p=1}^{P}(\pi_{n,p} - \pi_{n\prime,p})^2}, \tag{5}$$

with $\pi_{n,p}$ and $\pi_{n\prime,p}$ the vector of the autoregressive coefficients for the n-th and $n\prime$-th stocks, respectively. If $P_1 \neq P_2$, we take $P = \max(P_1, P_2)$ and $\pi_{n,p} = 0$ for $P > P_1$ and, similarly, $\pi_{n\prime,p} = 0$ for $P > P_2$.

All these measures ignore a crucial stylized fact: the time-varying nature of volatility. As proposed by many authors (Otranto 2008; Caiado and Crato 2010; D'Urso et al. 2013), if

we aim to cluster time series with similar volatility behaviour, we should consider a distance of model-based type between estimated parameters of GARCH processes. The standard GARCH(p,q) model of Bollerslev (1986) can be specified as

$$r_t - \mu_t = \epsilon_t$$
$$\epsilon_t = \sigma_t z_t \quad \text{with} \quad z_t \sim \mathcal{N}(0,1),$$

where z_t is called *innovation* and it is a process with zero mean and unit variance, while σ_t is a univariate stochastic process independent from z_t of the form

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-p}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-q}^2,$$

with $\omega > 0$, $0 \leq \alpha_i < 1$, $0 \leq \beta_j < 1$ and $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$. Parameters' estimation can be easily done by maximum likelihood. According to Caiado and Crato (2010); D'Urso et al. (2013), assuming a GARCH(1,1) process, the estimated parameters $\hat{\alpha}$ and $\hat{\beta}$ for each i-th time series can be stored into a matrix $\mathbf{T} = (\hat{\alpha}, \hat{\beta})$ with Ω the covariance matrix associated to the estimates contained in the $\mathbf{T}$'s. Therefore, we can consider the following Mahalanobis-like distance between two returns time series $r_{n,t}$ and $r_{n',t}$

$$d_{\text{GARCH}_{n,n'}} = \sqrt{(\mathbf{T}_n - \mathbf{T}_{n'})' \, \Omega^{-1} \, (\mathbf{T}_n - \mathbf{T}_{n'})}, \tag{6}$$

where through the weighting matrix inverse Ω^{-1} we also account for the uncertainty in the parameter estimation step.

In the end, it is worthy to mention a strand of literature clustering time series based on their distributional characteristics. This is particularly important given the well-known relevance of returns distribution in finance (Dhesi et al. 2021; Jondeau and Rockinger 2012). The idea of clustering time series based on their distributional characteristics is originally due to Nanopoulos et al. (2001). Successively, Wang et al. (2006), and Fulcher and Jones (2014) proposed approaches of clustering based on multiple features, including static mean, variance, skewness and kurtosis. Bastos and Caiado (2021) adopted such features for clustering financial time series. Recently, authors proposed the use of time-varying distribution features for clustering financial time series (e.g. see Cerqueti et al. 2022)

The studies mentioned above do not assume any underlying probability distribution for the time series, but use sample estimators for those features. However, model-based approaches for clustering require the definition of a probability distribution able to describe all the time series in the sample, which differentiates in terms of estimated parameters (e.g. see Mattera et al. 2021). A critical desired property of these approaches is that maximum likelihood estimation of the parameters is possible.

In presence of a general $p(\cdot)$ density, the distribution-based clustering approaches consider the following $(N \times J)$ matrix F

$$\mathbf{F} = \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,J} \\ \vdots & \vdots & \vdots & \vdots \\ f_{n,1} & f_{n,2} & \cdots & f_{n,J} \\ \vdots & \vdots & \vdots & \vdots \\ f_{N,1} & f_{N,2} & \cdots & f_{N,J} \end{bmatrix}, \tag{7}$$

where on the J columns of $\mathbf{F}$ there are the $j = 1, \ldots, J$ parameters for the N assets that are indexed by the rows. Clearly, in specifying the density $p(\cdot)$, it would be advantageous to choose a very general distribution to account for a wide range of possible exceptional cases. The observed characteristics of financial time series motivated the exploration of distributions that can accommodate properties such as fat-tailedness and skewness. In finance, a commonly employed distribution is the Skewed Exponential Power Distribution (SEPD) Fernandez et al. (1995); Fernández and Steel (1998); Theodossiou (2015); Komunjer (2007), that generalizes the Exponential Power Distribution for skewness. The SEPD is characterized by 4 parameters, i.e. location μ, scale ϕ, skewness λ and shape v. A random variable is said to have a Skewed Exponential Power Distribution if its probability density function is the following (Ayebo and Kozubowski 2003):

$$p(y; \mu, \sigma, v, \lambda) = \frac{v}{\sigma \Gamma\left(1 + \frac{1}{v}\right)} \frac{\lambda}{1 + \lambda^2}$$
$$\exp\left(-\frac{\lambda^p}{\sigma^v} [(z - \mu)^+]^v - \frac{1}{\sigma^v \lambda^v} [(z - \mu)^-]^v \right), \tag{8}$$

where:

$$(z - \mu)^+ = \max(z - \mu; 0) \quad \text{and}$$
$$(z - \mu)^- = \max(\mu - z; 0).$$

In literature seemingly different classes of SEPD distribution have been constructed (Ayebo and Kozubowski 2003; Theodossiou 2015). However, as suggested by Zhu and Zinde-Walsh (2009), all of them are reparametrizations of the SEPD proposed by Fernandez et al. (1995); Fernández and Steel (1998). An essential feature of the EPD is that it includes many common distributions as special cases, depending on the value of shape v and skewness λ parameters (see Fig. 1).

In particular, for $\lambda = 1$, the distribution is symmetric about μ so we obtain the symmetric exponential power distribution. If $\lambda = 1$ and $v = 2$ we obtain the Gaussian distribution. In the case $\lambda \neq 1$, by letting $v = 1$ we obtain the skewed Laplace distribution. Moreover, for $v = 2$ and $\lambda \neq 1$, we obtain

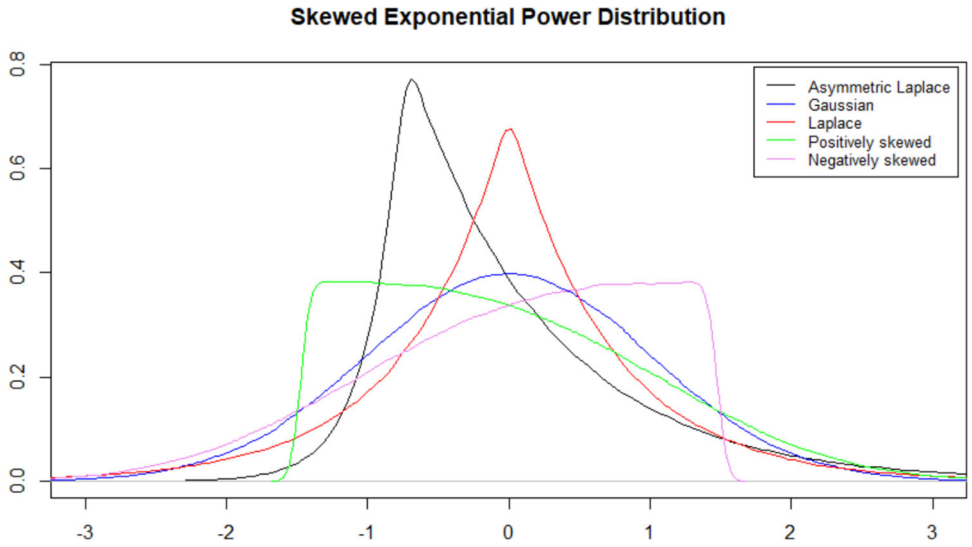

Fig. 1 Skewed Exponential Power Distribution for different values of shape and skewness

the skewed normal distribution as defined in Mudholkar and Hutson (2000). Many financial applications of the EPD as well as its skewed extensions have been considered (Nelson 1991; Cerqueti et al. 2020).

The flexibility of the SEPD can be successfully exploited in the clustering process if the aim is to form distribution-based clusters. In the case which the time series that are generated by a Skewed Exponential Power Distribution of parameters μ_n, σ_n, p_n and λ_n, the matrix

$$\mathbf{F}_{\text{sepd}} = \begin{bmatrix} \mu_1 & \sigma_1 & p_1 & \lambda_1 \\ \mu_2 & \sigma_2 & p_2 & \lambda_2 \\ \vdots & \vdots & \vdots & \vdots \\ \mu_n & \sigma_n & p_n & \lambda_n \\ \vdots & \vdots & \vdots & \vdots \\ \mu_N & \sigma_N & p_N & \lambda_N \end{bmatrix} \tag{9}$$

becomes the input for measuring the proximity of the time series. In the case of a generic distribution-based clustering with static parameters, a reasonable dissimilarity is given by the Euclidean distance between parameters

$$d_{n,n'} = \sqrt{\sum_{j=1}^{J} (\mathbf{F}_n - \mathbf{F}_{n'})^2}, \tag{10}$$

where $\mathbf{F}_n$ represents the n-the row of the matrix (7).

3 The clustered portfolio investment strategy

Most of the applications to financial time series clustering have been based on either the hierarchical (e.g. see Mantegna 1999; Caiado and Crato 2010; Raffinot 2017) or partitional algorithms (e.g. see Nanda et al. 2010; D'Urso et al. 2016; Lorenzo and Arroyo 2023).

Hierarchical clustering methods work by grouping time series into a tree of clusters. However, the performance of hierarchical clustering methods often suffers from their inability to adjust once a merge decision has been executed (Liu et al. 2021). Moreover, the hierarchical algorithms typically are time-consuming (Xie et al. 2020) with quadratic complexity, while for the partitional algorithms like the k-means it is linear. Therefore, despite hierarchical approaches representing a widely used alternative for clustering financial time series, they are not recommended for building portfolios in a high-dimensional setting, where the number of assets is relatively large.

The k-means algorithm is computationally less challenging. For this reason, the k-Means algorithm is one of the most popular clustering approaches aiming to partition the time series into a predetermined number of clusters K. The main drawback of partitional algorithms is that the number of clusters has to be identified in advance. Several approaches can be used to this aim. However, following many other authoritative studies (Arbelaitz et al. 2013; Batool and Hennig 2021), this issue can be overcome by choosing the number of clusters through meaningful criteria, such as the Average Silhouette Width (ASW). The k-means clustering algorithm relies on an iterative scheme based on the minimization of an objective function, which is usually chosen to be the total distance between all patterns from their respective cluster prototype

$$\min_{\{\mathcal{C}_k\}} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} d_{i,k}^2 \tag{11}$$

where N is the number of the time series to be grouped, $\mathcal{C}_k$ is the k-th cluster of size N_k, K is the number of clusters (a priori fixed), k represents the centre such that $d_{i,k}$ is the distance between each time series i from the prototype of the k-th cluster, that is equal to

$$r_{k,t} = \frac{1}{N_k} \sum_{i \in \mathcal{C}_k} r_{i,t} \tag{12}$$

in the case of the standard Euclidean distance. Similar to k-means, k-medoids algorithms are also usually considered for building portfolios. The k-medoids algorithm belongs, like the k-means, to the class of partitional approaches but, differently from the latter, provides a timid robustification (Garcia-Escudero and Gordaliza 2005). With the k-medoids approach, the prototypes of each group are real time series belonging to the sample, instead of averages, as happens with the k-means algorithm. The possibility of obtaining non-fictitious representative time series in the clusters is very appealing and helpful in many application domains, and it can improve the interpretability of the clusters.

In the context of finance, however, the results obtained with the k-medoids turn out to be less easy to interpret, at least in terms of financial portfolio theory. The prototype of the k-means algorithm has a clear financial interpretation: each k-th prototype is the equally weighted portfolio of all the assets included in the k-th clusters (12). This important distinguishing financial interpretation is not possible to obtain with k-medoids approaches. Therefore, while in the k-medoids investing in the prototype means investing in a single asset, with the k-means we invest in a diversified fund.

Beyond these classical methods, other unsupervised learning techniques could be considered, such as DBSCAN or GMM clustering approaches. On one hand, DBSCAN, is particularly effective for detecting clusters of varying shapes and densities, making it useful in non-Euclidean feature spaces or when the data exhibits noise. However, its applicability in the context of financial portfolio construction is limited, as DBSCAN does not naturally provide prototypes that can be interpreted as portfolios. GMM clustering, on the other hand, makes distributional assumption on the time series and/or their statistical features, making it in general less adopted for clustering financial time series.

In practice, following k-means approach, clustered portfolios can be formed by the application of any diversification rule to the $K \geq 2$ subsets of assets. The first step of the clustering-based investment requires the selection of the clustering approach, that is the employed proximity definition. Once a partition into K groups is found, K equally weighted clustered portfolios can be obtained based on the stocks belonging to each of the $k \in K$ clusters. In the end, the optimal amount of wealth associated to each of the $K << N$ funds can be defined according to any kind of optimization strategy. A simple approach consists of forming K well-diversified portfolios by applying optimal diversification across the K funds. This strategy turns out to provide a scaling to the equally weighted to each k-th fund constructed with clustering. At the end of the process, by clustered portfolio strategy one invests in all the N stocks in the initial sample.

We notice, however, that in principle any diversification rule across the K equally weighted portfolios can be adopted. To optimally choose w_k, for example, we can use the minimum-variance diversification rule

$$\begin{aligned} &\min_{w} \mathbf{w}' \boldsymbol{\Sigma} \mathbf{w}, \\ &\text{s.t.} \quad \mathbf{w}' \iota = 1 \end{aligned} \tag{13}$$

where $\boldsymbol{\Sigma}$ is the K funds covariance matrix, ι the K-dimensional vector of ones and $\mathbf{w} = [w_1, \ldots, w_k, \ldots, w_K]'$ is the vector of K weights associated to each k-th clustered portfolios. The well-known optimal solution is given by

$$\mathbf{w}^* = \frac{\boldsymbol{\Sigma}^{-1} \iota}{\iota' \boldsymbol{\Sigma}^{-1} \iota}. \tag{14}$$

Another well-known solution is to diversify by maximizing the investor's mean-variance utility function

$$\max_{\mathbf{w}} \mathbf{w}' \boldsymbol{\mu} - \gamma \mathbf{w}' \boldsymbol{\Sigma} \mathbf{w}, \tag{15}$$

with $\boldsymbol{\mu}$ the mean vector of the asset returns and γ the risk aversion coefficient. The optimal solution is given by

$$\mathbf{w}^* = \gamma^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}. \tag{16}$$

We stress that in a high-dimensional setting the covariance matrix $\boldsymbol{\Sigma}$ cannot be inverted, while with increasing assets N and fixed T it becomes increasingly ill-conditioned. Therefore, clustered portfolios represent an easy-to-implement solution to deal with such settings. Alternative approaches, based on particular covariance estimators (e.g. see Fan et al. 2013; Ledoit and Wolf 2017) need to be adopted. In this paper, we show that clustered portfolios can represent a better alternative to such approaches.

4 Empirical analysis

4.1 Data and experimental set-up

We study the performance of clustered portfolio strategy in both standard and high-dimensional scenarios. For this aim, two different datasets are considered. In the first one, we consider the monthly time series of stocks included in the Russell

2000 Index. For the same period, in the second case, we consider the monthly time series of stocks traded in the Dow Jones Industrial Average index from January 2003 to December 2023. We exclude stocks with missing values. Therefore, the final sample size in the first dataset is $N = 556$ stocks, while in the second case we have $N = 28$ as shown in Fig. 2. In both cases, we observe $T = 252$. The use of monthly data is justified by the fact that daily returns are much noisy and that medium-long run allocation strategies are usually constructed considering low-frequency time series.

To study the performances of clustered portfolios in out-of-sample, we use the rolling-window strategy described in DeMiguel et al. (2009). Given N time series of returns observed for T months, we choose an estimation widow equal to M to form K clusters considering alternative distance definitions as described in Sect. 2. From this we estimate the K clusters to form K naive equally weighted initial portfolios. Then, we estimate the covariance structure across the K funds, which is needed for the use of both global minimum-variance and mean-variance diversification rules. To estimate the covariance structure we use the static sample covariance estimator. This process is iteratively repeated by adding the return for the next period in the dataset and dropping the earliest one until the end of the dataset is reached. The result is, therefore, a time series of length $(T - M)$ of portfolio returns.

Usually, a common choice is $M = 120$ for monthly data (DeMiguel et al. 2009). However, the clustering algorithms perform differently for long and short time series (Díaz and Vilar 2010). Conversely, it can be the case that clustering approaches that work well with short time series perform instead poorly with long time series and vice versa. Since M represents the time series length within each iteration, we therefore also consider $M = 180$. In the first case, we assume 10 years of data are used for clustering and parameter estimation, while in the second case, 15 years are considered.

Given the time series of monthly out-of-sample portfolio returns, we then compute the out-of-sample Sharpe ratio of the portfolio obtained using the j-th strategy, SR_j, defined as the sample mean of out-of-sample portfolio returns divided by its standard deviation:

$$\text{Sharpe}_j = \frac{\hat{\mu}_j}{\hat{\sigma}_j} \tag{17}$$

where $\hat{\mu}_j$ is the average of the $(T - M)$ out of sample returns for the portfolio using the j-th clustering approach and $\hat{\sigma}_j$ its standard deviation. The investment strategies based on clustering are compared with the standard approaches where clustering is not involved.

The Sortino ratio and modified Sharpe ratio are, however, commonly favoured by practitioners because they address specific aspects of risk that the traditional Sharpe ratio may not fully capture. It could be interesting to weigh more the risk associated with negative returns risk. Investors are typically more concerned about losses than gains, especially those with a lower risk tolerance. Moreover, the standard deviation used in the Sharpe ratio could be a biased measure of risk, and other risk measures could be adopted.

The Sortino ratio is a variation of the Sharpe ratio that emphasizes downside risk, specifically the volatility of negative returns. It measures the excess return per unit of downside volatility. The downside volatility is calculated as the standard deviation of negative returns. The Sortino ratio is given by

$$\text{Sortino}_j = \frac{\hat{\mu}_j}{\hat{\sigma}_j^-}, \tag{18}$$

where $\hat{\sigma}_j^-$ is the downside risk, that is the standard deviation of negative returns.

The modified Sharpe ratio, also known as the VaR-Sharpe ratio, extends the concept of the Sharpe ratio by incorporating Value at Risk (VaR) in the denominator. The VaR is a measure of the maximum potential loss of an investment over a specified time horizon at a given confidence level. The VaR-Sharpe ratio measures the excess return per unit of VaR, i.e.

$$\text{VaR-Sharpe}_j = \frac{\hat{\mu}_j}{\text{VaR}_j}. \tag{19}$$

In the case of a high-dimensional experiment where $N > T$, the investment strategies without clustering are implemented using Ledoit and Wolf (2003) covariance estimator. In the low-dimensional setting, instead, the sample covariance estimator can be considered. Table 1 summarizes the investment strategies compared empirically.

4.2 Clustered portfolio analysis

In this section, we delve into a comprehensive analysis of portfolio performance across different dimensions and scenarios. We evaluate various strategies under both high-dimensional and low-dimensional portfolio settings, considering different time series lengths and risk measures. The portfolio strategies are assessed using key performance metrics such as the Sharpe ratio, Sortino ratio, and VaR-Sharpe ratio under both the GMV (Global Minimum Variance) rule and MV (Mean-Variance) rule approaches.

Through this detailed examination, we aim to provide insights into the effectiveness of each strategy in maximizing risk-adjusted returns while effectively managing downside risk and volatility. The analysis sheds light on the resilience of portfolio optimization techniques in dynamic market environments, offering valuable guidance for investors and

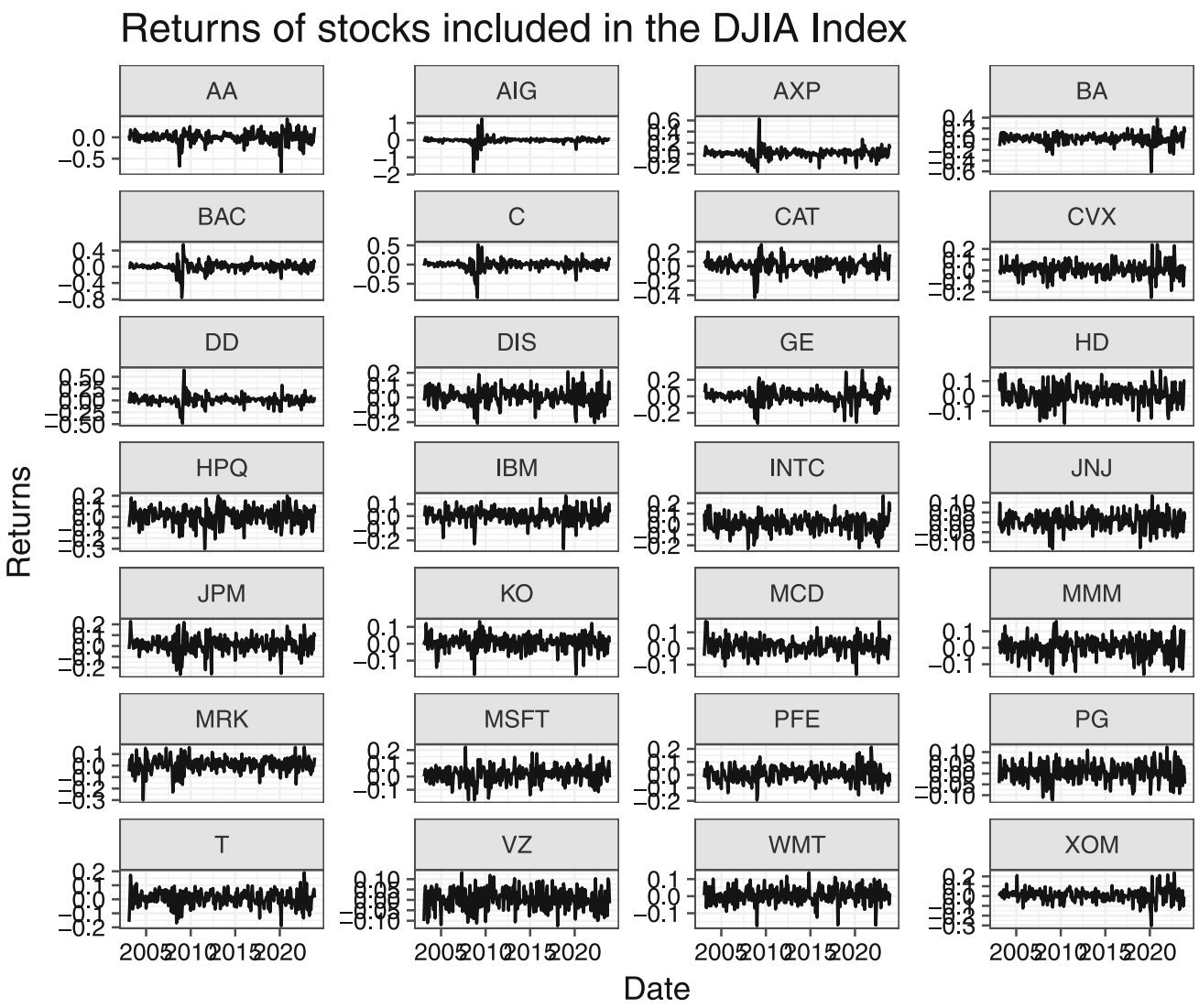

Fig. 2 Returns of the selected stocks included in the DJIA Index

Table 1 Implemented strategies, both global minimum variance and mean-variance diversification rules are considered

Symbol	Description
Benchmark	
w_{SC}	Investment strategy with sample covariance plug-in
w_{LW}	Investment strategy with Ledoit and Wolf (2003) plug-in
Clustered portfolios	
w_{EUCL}	Investment strategy with Euclidean distance
w_{COR}	Investment strategy with correlation-based distance
w_{NPER}	Investment strategy with peridiogram-based distance
w_{HURST}	Investment strategy with Hurst exponent-based distance
w_{ARMA}	Investment strategy with ARMA-based distance
w_{GARCH}	Investment strategy with GARCH-based distance
w_{SEPD}	Investment strategy with t-student distribution distance

practitioners seeking robust strategies for portfolio management.

4.2.1 High-dimensional portfolio

In this subsection, we focus on the analysis of portfolio performance in high-dimensional settings. We explore the effectiveness of various strategies under scenarios characterized by a large number of assets (N) and different time series lengths (M). The evaluation is conducted using a range of risk measures, including the Sharpe ratio, Sortino ratio, and VaR-Sharpe ratio, under both the GMV (Global Minimum Variance) rule and MV (Mean-Variance) rule approaches. Table 2 shows the out-of-sample results considering an estimation window equal to $M = 120$.

GMV Rule. Under the GMV rule approach, SEPD consistently outperformed other models across all risk measures. This suggests that SEPD not only achieves higher risk-adjusted returns but also effectively manages downside risk and volatility. For instance, SEPD achieved a Sharpe ratio of 0.2775, significantly higher than the next best-performing model, indicating its ability to generate superior returns per unit of risk. Similarly, SEPD demonstrated a Sortino ratio of 0.4905, reflecting its efficiency in delivering positive returns relative to downside volatility. Additionally, SEPD showcased a VaR-Sharpe ratio of 0.2214, highlighting its robustness in managing both volatility and downside risk simultaneously (Table 2).

MV Rule. Similarly, under the MV rule approach, SEPD emerged as the top-performing model across all risk measures. The robust performance of SEPD in this approach reaffirms its effectiveness in maximizing returns while mitigating risk. For example, SEPD achieved a Sharpe ratio of 0.4905, surpassing all other models, indicating its superior risk-adjusted returns. Additionally, SEPD demonstrated a Sortino ratio of 0.2214, indicating its ability to generate higher returns relative to downside risk compared to alternative models. Furthermore, SEPD exhibited a VaR-Sharpe ratio of 0.2194, highlighting its effectiveness in managing volatility while considering downside risk.

The consistent outperformance of SEPD across all risk measures underscores its effectiveness in managing estimation error and high-dimensional data. This suggests that SEPD offers a robust solution for portfolio optimization in challenging environments, providing investors with strong risk-adjusted returns while effectively managing volatility and downside risk. The substantial performance gap between SEPD and other models across all risk measures highlights its superiority in generating superior risk-adjusted returns and underscores its resilience in high-dimensional settings.

Table 3 shows the out-of-sample results by increasing the estimation window to $M = 180$.

GMV Rule. In the extended scenario with a longer time series length (M=180), "PER" emerged as the top-performing model based on the Sharpe ratio. However, SEPD maintained its competitive edge across all risk measures, including the Sortino and VaR-Sharpe ratios, highlighting its robustness in managing volatility and downside risk. Despite the slight shift in performance, SEPD continues to demonstrate its effectiveness in delivering superior risk-adjusted returns.

MV rule. Similarly, under the MV rule approach in Table 3, "PER" demonstrated strong performance based on the Sharpe ratio. Nevertheless, SEPD remained a robust contender across all risk measures, indicating its effectiveness in managing downside risk while delivering strong risk-adjusted returns. The consistent performance of SEPD reaffirms its suitability for robust portfolio optimization, providing investors with stability and reliability in dynamic market conditions.

The performance trends observed in Table 3 reinforce the resilience of SEPD in managing estimation error and high-dimensional data. It consistently outperforms all risk measures, reaffirming its suitability for robust portfolio optimization. The differences across the rows highlight the varying performance levels of different models, with SEPD consistently outperforming other models in terms of risk-adjusted returns and demonstrating its adaptability in dynamic market conditions.

4.2.2 Low-dimensional portfolio

The results from Tables 4 and 5 provide insights into the portfolio performance of various strategies in low-dimensional scenarios, considering different risk measures and approaches.

In Table 4, which presents the analysis for a scenario with a time series length of $M = 120$ and $N = 28$, the strategies are evaluated under both the GMV (Global Minimum Variance) rule and MV (Mean-Variance) rule approaches. The performance metrics include the Sharpe ratio, Sortino ratio, and VaR-Sharpe ratio. Notably, the SC strategy emerges as the top-performing model under the GMV rule approach, exhibiting the highest Sharpe, Sortino, and VaR-Sharpe ratios among the strategies considered. This indicates its effectiveness in maximizing risk-adjusted returns while managing downside risk and volatility.

Under the MV rule approach in Table 4, the SC strategy also demonstrates strong performance across all risk measures, further reinforcing its suitability for portfolio optimization in low-dimensional scenarios.

Moving to Table 5, which depicts the results for a scenario with an extended time series length of $M = 180$ and $N = 28$, similar trends are observed. The SC strategy maintains its competitive edge, emerging as the top-performing model under both the GMV rule and MV rule approaches.

Table 2 Portfolio performance of the strategies considered in Table 1: different measures, $M = 120$ and $N = 556$

Approaches	GMV rule			MV rule		
	Sharpe	Sortino	VaR-Sharpe	Sharpe	Sortino	VaR-Sharpe
SC	–	–	–	–	–	–
LW	0.1639	0.2444	0.1166	**0.1886**	0.3048	0.1203
EUCL	0.1874	0.3083	0.1334	0.0399	0.0797	0.0400
COR	0.0744	0.1062	0.0543	0.0025	0.0036	0.0039
PER	0.2183	0.3553	0.1649	−0.0972	−0.0980	−0.1653
HURST	0.1322	0.1918	0.0927	0.0956	0.3792	0.3695
ARMA	0.0454	0.0591	0.0297	0.1685	0.4003	0.4055
GARCH	0.1046	0.1478	0.0649	−0.0866	−0.0877	−0.5272
SEPD	**0.2775**	**0.4905**	**0.2214**	0.1501	**0.4479**	**0.2194**

The best model is highlighted with the bold font

Table 3 Portfolio performance of the strategies considered in Table 1: different measures, $M = 180$ and $N = 556$

Approaches	GMV rule			MV rule		
	Sharpe	Sortino	VaR-Sharpe	Sharpe	Sortino	VaR-Sharpe
SC	–	–	–	–	–	–
LW	0.0017	0.0023	0.0010	0.1692	0.2688	0.1214
EUCL	0.0939	0.1427	0.0681	0.1256	0.2068	0.0870
COR	−0.0633	−0.0788	−0.0363	0.0742	0.1302	0.0674
PER	0.1455	0.2259	0.1100	0.2073	0.3518	0.1513
HURST	0.0439	0.0616	0.0307	0.1302	0.2399	0.1154
ARMA	0.0631	0.0864	0.0391	0.0769	0.1274	0.0592
GARCH	0.0150	0.0204	0.0085	0.0592	0.0823	0.0323
SEPD	**0.2071**	**0.3456**	**0.1610**	**0.2515**	**0.4731**	**0.1775**

The best model is highlighted with the bold font

Table 4 Portfolio performance of the strategies considered in Table 1: different measures, $M = 120$ and $N = 28$

Approaches	GMV rule			MV rule		
	Sharpe	Sortino	VaR-Sharpe	Sharpe	Sortino	VaR-Sharpe
SC	**0.2350**	**0.3880**	**0.1706**	**0.2048**	**0.3453**	0.1405
LW	0.2243	0.3635	0.1479	0.1982	0.3325	**0.1537**
EUCL	0.1929	0.3036	0.1152	0.1111	0.8841	0.3936
COR	0.1819	0.2828	0.1253	0.0437	0.0570	0.0305
PER	0.2162	0.3414	0.1395	0.1214	0.4051	0.2119
HURST	0.1585	0.2382	0.1093	0.0521	0.1020	0.1475
ARMA	0.1682	0.2420	0.1174	0.1609	0.3439	0.1953
GARCH	0.2104	0.3379	0.1398	−0.0482	−0.0530	−0.3568
SEPD	0.2214	0.3503	0.1440	0.1948	0.3288	0.1412

The best model is highlighted with the bold font

Once again, it exhibits superior risk-adjusted returns and effective management of downside risk and volatility compared to other strategies.

These findings underscore the robustness of the SC strategy in low-dimensional portfolio settings, highlighting its ability to deliver strong risk-adjusted returns while effectively navigating market dynamics. The consistency of its performance across different risk measures and approaches reaffirms its suitability for investors seeking stability and reliability in portfolio optimization strategies.

Overall, the results presented in Tables 4 and 5 provide valuable insights for practitioners and investors, guiding their decision-making processes in low-dimensional portfolio management.

Table 5 Portfolio performance of the strategies considered in Table 1: different measures, $M = 180$ and $N = 28$

Approaches	GMV rule			MV rule		
	Sharpe	Sortino	VaR-Sharpe	Sharpe	Sortino	VaR-Sharpe
SC	0.1566	**0.2538**	**0.0976**	0.1364	0.2357	**0.1206**
LW	0.1485	0.2372	0.0984	0.1403	0.2364	0.1062
EUCL	0.1000	0.1489	0.0647	0.1120	0.1693	0.0735
COR	0.1017	0.1522	0.0644	0.1075	0.1637	0.0701
PER	0.0989	0.1492	0.0611	0.1108	0.1669	0.0686
HURST	0.1093	0.1647	0.0792	0.0979	0.1467	0.0703
ARMA	0.1319	0.2034	0.0929	0.0508	0.0766	0.0332
GARCH	0.1502	0.2289	0.0917	0.1513	0.2289	0.0858
SEPD	**0.1606**	0.2491	0.0935	**0.1908**	**0.2987**	0.1109

The best model is highlighted with the bold font

5 Conclusions

The application of time series clustering techniques in portfolio construction presents a promising avenue for enhancing investment outcomes in financial markets. Clustering, a fundamental algorithm in data mining, finds extensive utility in exploratory analysis, anomaly detection, and classification tasks. However, its adaptation to time series data introduces additional complexities, particularly in the domain of finance where empirical regularities, commonly known as stylized facts, must be carefully considered.

In our investigation, we observed that time series clustering techniques play a pivotal role in high-dimensional portfolio analysis, where traditional optimization strategies face challenges in estimating covariance matrices accurately. Notably, the k-means clustering algorithm emerges as a powerful tool for identifying smaller sets of stocks to construct diversified funds, with prototype portfolios resembling equally weighted allocations. Despite the effectiveness of equally weighted portfolios in high-dimensional settings, optimal diversification strategies such as mean-variance and GMV portfolios offer refined approaches to portfolio construction by optimizing the risk-return trade-off.

In the paper, we consider the curse of dimensionality from two perspectives. First, we reduce the set of N stocks to N_k, $\forall k = 1, \ldots, K$ and such that $N_k << N$, thus making the portfolio selection choice feasible in large dimension. Second, by clustering time series considering relevant say P features rather than T temporal ordinates, we reduce the features for clustering from T to P, with $P << T$.

The results presented in the paper illustrate the portfolio performance of various strategies under different risk measures and approaches. The analysis focuses on both high-dimensional and low-dimensional portfolios, considering different scenarios with varying time series lengths.

In the high-dimensional portfolio analysis, the strategies are evaluated under the GMV (Global Minimum Variance) rule and MV (Mean-Variance) rule approaches. The performance metrics include the Sharpe ratio, Sortino ratio, and VaR-Sharpe ratio. Notably, the SEPD strategy consistently outperforms other models across all risk measures under both the GMV and MV rule approaches. This indicates that SEPD not only achieves higher risk-adjusted returns but also effectively manages downside risk and volatility. Under the GMV rule approach, SEPD demonstrates superior performance compared to other models, as evidenced by its higher Sharpe, Sortino, and VaR-Sharpe ratios. Similarly, under the MV rule approach, SEPD emerges as the top-performing model across all risk measures, showcasing its effectiveness in maximizing returns while mitigating risk. The consistent outperformance of SEPD underscores its effectiveness in managing estimation error and high-dimensional data, making it a robust solution for portfolio optimization in challenging environments. The substantial performance gap between SEPD and other models across all risk measures highlights its superiority in generating superior risk-adjusted returns and underscores its resilience in high-dimensional settings.

In the low-dimensional portfolio analysis, similar trends are observed, with SEPD consistently performing well across all risk measures and approaches. Despite variations in the performance of other models, SEPD maintains its competitive edge, reaffirming its suitability for robust portfolio optimization.

Our comprehensive comparison of various time series clustering techniques in the context of portfolio performance revealed valuable insights. Specifically, the SEPD strategy consistently outperformed other models across different risk measures and portfolio dimensions, demonstrating its robustness in managing estimation error and high-dimensional data. This underscores the superiority of SEPD in generating superior risk-adjusted returns while effectively managing downside risk and volatility, making it a compelling solution for portfolio optimization in challenging environments.

In conclusion, leveraging the advantages of time series clustering techniques in portfolio construction holds sig-

nificant potential for investors and portfolio managers. By carefully selecting appropriate clustering methods and considering the relationship between clustering techniques and portfolio outcomes, investors can enhance their investment strategies and achieve superior risk-adjusted returns in dynamic market conditions. Our analysis contributes valuable insights to the field, guiding practitioners towards more effective portfolio management strategies in the ever-evolving landscape of financial markets.

While our study highlights the effectiveness of clustering approaches, future research could explore promising alternative methods for portfolio construction, such as factor-based models. These approaches, which focus on identifying and exploiting common risk factors driving asset returns, offer a complementary framework to clustering as they do not rely on a similarity criterion among stocks.

A particularly intriguing avenue for further investigation is the integration of clustering techniques within factor investing strategies (e.g. see De Nard et al. 2021). For instance, clustering could be employed to refine factor portfolios by grouping assets based on shared factors, thereby enhancing the robustness of factor models. Some studies (e.g. see Ando and Bai 2017) have begun to explore this synergy, demonstrating the usefulness of clustering in improving the performances of factor models in explaining and forecasting time series. Additionally, future research could investigate the use of more sophisticated distance measures, such as Kullback–Leibler (KL) divergence, to enhance clustering performance. While our study employs a distribution-based clustering approach that has shown strong performance, KL divergence could offer complementary insights by accounting for differences in probability distributions. Expanding the scope of research to include hybrid approaches that combine clustering with factor-based models and more advanced distance metrics could provide deeper insights, advancing the toolkit available for portfolio management.

Funding Open access funding provided by Università degli Studi della Campania Luigi Vanvitelli within the CRUI-CARE Agreement. This work (coauthor Parmjit Kaur) is partially supported by the project "A better understanding of socio-economic systems using quantitative methods from Physics "funded by European Union-Next generation EU and Romanian Government, under National Recovery and Resilience Plan for Romania, contract no.760034/ 23.05.2023, cod PNRR-C9-I8-CF 255/ 29.11.2022, through the Romanian Ministry of Research, Innovation and Digitalization, within Component 9, Investment I8.

Data availability Data and codes are available upon request.

Declarations

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of interest The authors have not disclosed any conflict of interest.

References

Ausloos M, Bronlet P (2003) Strategy for investments from Zipf law (s). Phys A 324(1–2):30–37

Ando T, Bai J (2017) Clustering huge number of financial time series: a panel data approach with high-dimensional predictors and factor structures. J Am Stat Assoc 112(519):1182–1198

Ausloos M, Cerqueti R, Lupi C (2017) Long-range properties and data validity for hydrogeological time series: the case of the Paglia river. Phys A 470:39–50

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recogn 46(1):243–256

Ayebo A, Kozubowski TJ (2003) An asymmetric generalization of Gaussian and Laplace laws. J Probab Stat Sci 1(2):187–210

Ausloos M, Lambiotte R (2007) Clusters or networks of economies? a macroeconomy study through gross domestic product. Phys A 382(1):16–21

Ausloos M (2000) Statistical physics in foreign exchange currency and stock markets. Phys A 285(1–2):48–65

Bastos JA, Caiado J (2021) On the classification of financial data with domain agnostic features. Int J Approx Reason 138:1–11

Batool F, Hennig C (2021) Clustering with the average silhouette width. Comput Stat Data Anal 158:107190

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Economet 31(3):307–327

Caiado J, Crato N (2010) Identifying common dynamic features in stock returns. Quant Finance 10(7):797–807

Caiado J, Crato N, Peña D (2006) A periodogram-based metric for time series classification. Comput Stat Data Anal 50(10):2668–2684

Cerqueti R, D'Urso P, De Giovanni L, Giacalone M, Mattera R (2022) Weighted score-driven fuzzy clustering of time series with a financial application. Expert Syst Appl 198:116752

Cerqueti R, Fanelli V (2021) Long memory and crude oil's price predictability. Ann Oper Res 299:895–906

Cerqueti R, Giacalone M, Mattera R (2020) Skewed non-Gaussian GARCH models for cryptocurrencies volatility modelling. Inf Sci 527:1–26

Cerqueti R, Mattera R (2023) Fuzzy clustering of time series with time-varying memory. Int J Approx Reason 153:193–218

D'Urso P, Cappelli C, Di Lallo D, Massari R (2013) Clustering of financial time series. Phys A 392(9):2114–2129

D'Urso P, De Giovanni L, Massari R (2016) Garch-based robust clustering of time series. Fuzzy Sets Syst 305:1–28

D'Urso P, De Giovanni L, Massari R, D'Ecclesia RL, Maharaj EA (2020) Cepstral-based clustering of financial time series. Expert Syst Appl 161:113705

Dimitrova V, Fernández-Martínez M, Sánchez-Granero M, Trinidad Segovia J (2019) Some comments on bitcoin market (in) efficiency. PLoS ONE 14(7):0219243

DeMiguel V, Garlappi L, Uppal R (2009) Optimal versus Naive diversification: How inefficient is the 1/n portfolio strategy? Rev Financ Stud 22(5):1915–1953
De Nard G, Ledoit O, Wolf M (2021) Factor models for portfolio selection in large dimensions: the good, the better and the ugly. J Financ Economet 19(2):236–257
Di Sciorio F (2023) Clustering analysis on hurst dynamic. Studies of Applied Economics
Dhesi G, Shakeel B, Ausloos M (2021) Modelling and forecasting the kurtosis and returns distributions of financial markets: irrational fractional Brownian motion model approach. Ann Oper Res 299:1397–1410
Díaz SP, Vilar JA (2010) Comparing several parametric and nonparametric approaches to time series clustering: a simulation study. J Classif 27(3):333–362
Fulcher BD, Jones NS (2014) Highly comparative feature-based time-series classification. IEEE Trans Knowl Data Eng 26(12):3026–3037
Fan J, Liao Y, Mincheva M (2013) Large covariance estimation by thresholding principal orthogonal complements. J R Stat Soc Ser B (Stat Methodol) 75(4):603–680
Fernandez C, Osiewalski J, Steel MF (1995) Modeling and inference with υ-spherical distributions. J Am Stat Assoc 90(432):1331–1340
Fernández C, Steel MF (1998) On Bayesian modeling of fat tails and skewness. J Am Stat Assoc 93(441):359–371
Gligor M, Ausloos M (2007) Cluster structure of EU-15 countries derived from the correlation matrix analysis of macroeconomic index fluctuations. Eur Phys J B 57:139–146
Garcia-Escudero LA, Gordaliza A (2005) A proposal for robust curve clustering. J Classif 22(2):185–201
Gubu L, Rosadi D et al (2020) Robust mean-variance portfolio selection using cluster analysis: A comparison between Kamila and weighted k-mean clustering. Asian Econ Financ Rev 10(10):1169
Iorio C, Frasso G, D'Ambrosio A, Siciliano R (2018) A p-spline based clustering approach for portfolio selection. Expert Syst Appl 95:88–103
Jondeau E, Rockinger M (2012) On the importance of time variability in higher moments for asset allocation. J Financ Economet 10(1):84–123
Komunjer I (2007) Asymmetric power distribution: theory and applications to risk measurement. J Appl Economet 22(5):891–921
Lorenzo L, Arroyo J (2023) Online risk-based portfolio allocation on subsets of crypto assets applying a prototype-based clustering algorithm. Financ Innov 9(1):25
Lahmiri S (2016) Clustering of Casablanca stock market based on Hurst exponent estimates. Phys A 456:310–318
Liao TW (2005) Clustering of time series data-a survey. Pattern Recogn 38(11):1857–1874
Ledoit O, Wolf M (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J Empir Financ 10(5):603–621
Ledoit O, Wolf M (2017) Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. Rev Financ Stud 30(12):4349–4388
Liu N, Xu Z, Zeng X-J, Ren P (2021) An agglomerative hierarchical clustering algorithm for linear ordinal rankings. Inf Sci 557:170–193
Mahmoudi MR (2021) A computational technique to classify several fractional Brownian motion processes. Chaos Solitons Fractals 150:111152
Mattera R, Athanasopoulos G, Hyndman R (2024) Improving out-of-sample forecasts of stock price indexes with forecast reconciliation and clustering. Quant Finance 24(11):1641–1667
Mantegna RN (1999) Hierarchical structure in financial markets. Eur Phys J B-Condens Matter Complex Syst 11(1):193–197
Mattera R, Di Sciorio F (2021) Option pricing under multifractional process and long-range dependence. Fluct Noise Lett 20(01):2150008
Mattera R, Di Sciorio F, Trinidad-Segovia JE (2022) A composite index for measuring stock market inefficiency. Complexity 2022:1–13
Mattera R, Giacalone M, Gibert K (2021) Distribution-based entropy weighting clustering of skewed and heavy tailed time series. Symmetry 13(6):959
Mudholkar GS, Hutson AD (2000) The epsilon-skew-normal distribution for analyzing near-normal data. J Stat Plan Inference 83(2):291–309
Nanopoulos A, Alcock R, Manolopoulos Y (2001) Feature-based classification of time-series data. Int J Comput Res 10(3):49–61
Nelson DB (1991) Conditional heteroskedasticity in asset returns: a new approach. Econom J Econom Soc 59:347–370
Nanda S, Mahanty B, Tiwari M (2010) Clustering Indian stock market data for portfolio management. Expert Syst Appl 37(12):8793–8798
Otranto E (2008) Clustering heteroskedastic time series by model-based procedures. Comput Stat Data Anal 52(10):4685–4698
Piccolo D (1990) A distance measure for classifying ARIMA models. J Time Ser Anal 11(2):153–164
Raffinot T (2017) Hierarchical clustering-based asset allocation. J Portf Manag 44(2):89–99
Ramos-Requena JP, Trinidad-Segovia J, Sánchez-Granero M (2017) Introducing Hurst exponent in pair trading. Phys A 488:39–45
Sánchez MÁ, Trinidad JE, García J, Fernández M (2015) The effect of the underlying distribution in Hurst exponent estimation. PLoS ONE 10(5):0127824
Soleymani F, Vasighi M (2022) Efficient portfolio construction by means of CVaR and k-means++ clustering analysis: Evidence from the NYSE. Int J Finance Econ 27(3):3679–3693
Tumminello M, Di Matteo T, Aste T, Mantegna RN (2007) Correlation based networks of equity returns sampled at different time horizons. Eur Phys J B 55:209–217
Theodossiou P (2015) Skewed generalized error distribution of financial assets and option pricing. Multinatl Finance J 19(4):223–266
Tola V, Lillo F, Gallegati M, Mantegna RN (2008) Cluster analysis for portfolio optimization. J Econ Dyn Control 32(1):235–258
Vogl M (2023) Hurst exponent dynamics of s&p 500 returns: implications for market efficiency, long memory, multifractality and financial crises predictability by application of a nonlinear dynamics analysis framework. Chaos, Solitons Fractals 166:112884
Wang X, Smith K, Hyndman R (2006) Characteristic-based clustering for time series data. Data Min Knowl Disc 13(3):335–364
Xie W-B, Lee Y-L, Wang C, Chen D-B, Zhou T (2020) Hierarchical clustering supported by reciprocal nearest neighbors. Inf Sci 527:279–292
Zhu D, Zinde-Walsh V (2009) Properties and estimation of asymmetric exponential power distribution. Journal of econometrics 148(1):86–99