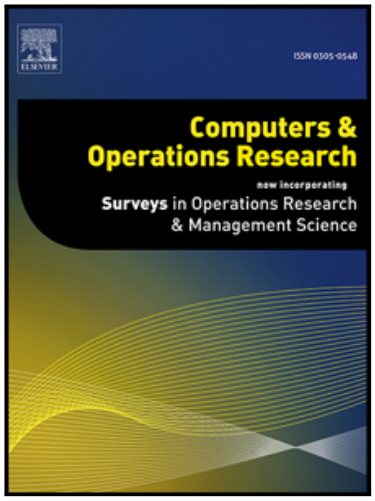
Journal Pre-proof



Clustering and portfolio selection problems: A unified framework

Justo Puerto, Moisés Rodríguez-Madrena, Andrea Scozzari

PII: S0305-0548(20)30008-3
DOI: https://doi.org/10.1016/j.cor.2020.104891
Reference: CAOR 104891

To appear in: *Computers and Operations Research*

Received date: 17 June 2019
Revised date: 27 November 2019
Accepted date: 13 January 2020

Please cite this article as: Justo Puerto, Moisés Rodríguez-Madrena, Andrea Scozzari, Clustering and portfolio selection problems: A unified framework, *Computers and Operations Research* (2020), doi: https://doi.org/10.1016/j.cor.2020.104891

Highlights

- We present a mixed-integer linear programming model for facing the clustering and asset selection problem in a unified framework
- We construct asset graphs based on stock price correlations in order to incorporate the dependency structure of the returns of each pair of stocks
- We use the p-median objective function in order to model the clustering phase in the MILP program
- The risk of a portfolio is evaluated through the CVaR risk measure
- We evaluate the effectiveness of our approach by providing an accurate out-of-sample performance on some real-world financial datasets

Clustering and portfolio selection problems: A unified framework

Justo Puerto[1,2], Moisés Rodríguez-Madrena [1,2], Andrea Scozzari[3]

[1] Institute of Mathematics University of Seville (IMUS), Seville, Spain
[2] Department of Statistics and OR, Universidad de Sevilla, 41012 Seville, Spain {puerto, madrena}@us.es
[3] Faculty of Economics, Università degli Studi Niccolò Cusano Roma, Italy andrea.scozzari@unicusano.it

January 14, 2020

Abstract

Given a set of assets and an investment capital, the classical portfolio selection problem consists in determining the amount of capital to be invested in each asset in order to build the most profitable portfolio. The portfolio optimization problem is naturally modeled as a mean-risk bi-criteria optimization problem where the mean rate of return of the portfolio must be maximized whereas a given risk measure must be minimized. Several mathematical programming models and techniques have been presented in the literature in order to efficiently solve the portfolio problem. A relatively recent promising line of research is to exploit clustering information of an assets network in order to develop new portfolio optimization paradigms. In this paper we endow the assets network with a metric based on correlation coefficients between assets' returns, and show how classical location problems on networks can be used for clustering assets. In particular, by adding a new criterion to the portfolio selection problem based on an objective function of a classical location problem, we are able to measure the effect of clustering on the selected assets with respect to the non-selected ones. Most papers dealing with clustering and portfolio selection models solve these problems in two distinct steps: cluster first and then selection. The innovative contribution of this paper is that we propose a Mixed-Integer Linear Programming formulation for dealing with this problem in a unified phase. The effectiveness of our approach is validated reporting some computational experiments on some real financial datasets.

Keywords: Portfolio selection; Clustering; p-median problem on networks; Conditional Value at Risk; Multicretiria optimization.

1 Introduction and motivation

Portfolio selection problem is among most the widely studied problems in financial literature since the seminal paper of Markowitz in 1952 [40]. He formalized the notion of diversification in investing and used the variance of asset prices as a proxy of risk. In spite of its success, the model proposed by Markowitz has received many criticisms, first and foremost due to the large estimation errors on the vector of expected returns and on the covariance matrix [15, 29, 41]. Hence, different directions have been proposed in the literature exploiting several strategies that consider improved estimates of correlation coefficients [22], alternative risk measures [1, 2], effective diversification approaches (see, e.g., [12, 20, 35] and the references therein). As a consequence, new mathematical programming models and techniques were proposed in the literature in order to efficiently solve the portfolio selection problems incorporating the above mentioned different strategies (the interested reader can refer to [37]).

At the early 2000s, some authors started in tackling the asset allocation problem via graph theory in which assets are identified by vertices of a complete graph G and distances d_{ij} assigned to pair of assets i and j incorporate the dependency structure of returns. Mantegna [38] was one of the first to construct asset graphs based on stock price correlations in order to detect the hierarchical organization inside a stock market. He showed that a minimal spanning tree of G provides an arrangement of stocks which selects the most relevant connections among them. Moreover, the minimal spanning tree provides, in a direct way, the subdominant ultrametric hierarchical organization of the assets in a market. Based on this idea, Onnela et al. [44] study the topology of stock networks that can be constructed by using return correlations. The aim is to disclose the relevant correlations from the correlation matrix. They start from an empty graph with no edges where the vertices correspond to assets. Then, one by one, they insert edges between the vertices according to the rank of their correlation strength, thus determining different stock graphs such as the one that minimally connects all the vertices (i.e., a spanning tree) until the entire graph being fully connected. For all these graphs, they study some relevant characteristics, such as topologically different growth types, number and size of clusters and clustering coefficient.

A relatively recent promising line of research is to exploit clustering information of an asset network in order to provide new portfolio optimization techniques. These approaches construct a portfolio by solving the classical Markowitz model by substituting the original correlation matrix with a correlation based clustering ultrametric matrix. The aim is twofold: (i) correlation based clustering may be seen as a filtering procedure and (ii) the portfolios selected by clustering algorithms are quite robust with respect to measurement noise due to the finiteness of sample size [7, 16, 17, 49]. In [49], the authors starts by determining the distances d_{ij} based on the Pearson Correlation measure among stocks. Then, they show that by applying the single linkage and the average linkage clustering procedure using the distances d_{ij}, they are able to obtain a new distance matrix $\hat{d_{ij}}$, where $\hat{d_{ij}}$ are the ultrametric distances between each pair of stocks. In particular, if a single linkage strategy is adopted, in contrast to the original distance matrix, the number of different element values in the ultrametric distance matrix cannot exceed $n-1$ (see [39]). Finally, they show that by solving the classical Markowitz model with the new ultrametric matrix can improve the reliability of the portfolio in terms of the ratio between predicted and realized risk. In [16] the authors construct the starting d_{ij} by using both the Pearson Correlation measure and two other measures, namely the Kendall rank correlation coefficient and the lower tail dependencies between each pair of assets.
To measure the level of interconnections of an asset with the whole system, in [16] the authors also use the clustering coefficient introduced by Watts and Strogatz in [50], which refers to the number of existing triangles around a vertex with respect to the number of potential ones. Finally, they obtain a positive definite matrix C of interconnections between pair of assets that is used in Markowitz's model in place of the original correlation matrix.

Other authors used graph-theoretic concepts to explore the global properties of stock markets by analyzing the structure of the underlying graph. In [8] the authors solve different classical NP-hard optimization problems to analyze the dependencies among stocks. A maximum clique problem was solved for detecting large clusters of similar and dissimilar stocks according to the natural criterion of pairwise correlation. Independent sets, which represent groups of vertices with no connections, are sought for in the stock graph to find well-diversified portfolios.

The common feature of the above strategies is that they are based on two different and independent phases: clustering first and then analyze the number and properties of the assets clustered this way. In fact, all these papers do not investigate how the clustering phase could be used as an effective tool in the portfolio selection process for assessing portfolios' optimal weights.

The main contribution of this paper is to present a clustering and portfolio selection strategy in a unified framework that, in principle, could be used and adapted for finding portfolios minimizing

different coherent risk measures usually considered in portfolio optimization. In this paper we model the clustering problem as a network p-median problem. This is a classical network facility location problem in the Operations Research literature. It was introduced in [30] and, given an n vertex graph G, the problem consists in finding a set of exactly p vertices (facilities) in order to minimize the sum of the distances (accessibility criterion) between the selected facilities and the other vertices of the graph. It is shown in [30] that the problem of finding a p-median of a network is NP-hard even when the network has a simple structure.

The p-median problem may be also interpreted in terms of cluster analysis, and the data clustering application is straightforward since it simply requires reinterpreting the medians as the median of a cluster instead of as a facility. In fact, in [27] the authors observed that when squared Euclidean distance is used, the problem becomes a discretized version of the well-known k-means problem in cluster analysis, and the optimal solution to the p-median problem results in a partition of the vertex set into homogeneous clusters. Ng and Han [43] showed that the p-median model is useful for detecting patterns and mining data as well as clustering. Thus, the p-median model can be a powerful tool for data mining applications. We also mention the research of Benati [3], Benati and García [4], and Benati et al. [5] on clustering based on p-median models. In particular, in [4] the authors outline that when the p-median is used as clustering algorithm, one has some advantages in terms of robustness and interpretation, for example, because the median representing the cluster is an element of the sample (see, also [31]).

In this paper we provide a novel mathematical programming formulation for the problem of simultaneously locating a set of p facilities in a network of stocks while selecting a subset of assets minimizing a given risk measure. In particular, here we consider the Conditional Value at Risk to measure the riskiness of an investment. We note that an asset selected as a facility can be considered as a *representative* of a cluster of stocks centered in it. We also show that our formulations are *portable* in the sense that there is a subset of constraints that remains unchanged regardless of the risk measure considered. We test our formulations on different real world financial datasets, and compare the results of the different formulations from a profitability point of view.

The paper is organized as follows. In Section 2 we introduce some necessary definitions and notation. In Section 3 we provide the mathematical programming formulation of our problem, in addition to a description on how to compute some parameters of the model. Section 4 is devoted to the computational experiments that show the effectiveness of our approach on some real financial datasets. Finally, in the concluding section some remarks and future research are outlined.

2 Notation, definitions and properties

Consider a finite, connected and undirected graph with no self-loops $G = (V, E)$. We denote by $V(G)$ and $E(G)$ the vertex and edge set, respectively, with $|V(G)| = n$ and $|E(G)| = m$. An edge $e \in E(G)$ is identified by a pair of vertices (i, j), with $i, j \in V(G)$. Suppose that a nonnegative real length $d(e) = d_{ij}$ is assigned to each edge $e \in E(G)$. For any pair of vertices $u, v \in V(G)$, we let d_{uv} denote the length of a shortest path in G connecting u and v. Consider a subset V_p of p vertices of $V(G)$; the distance $d_u(V_p)$ between a vertex u and V_p is defined as the smallest of the distances from u to the vertices in V_p, that is: $d_u(V_p) = \min\{d_{uv} | v \in V_p\}$. Thus, we define:

$$F(V_p) = \sum_{u \in V(G)} d_u(V_p),$$

the *sum of the distances* of all the vertices in G to the set V_p. The p-median problem on G consists in finding a set (of facilities) V_p^* of cardinality p such that

$$F(V_p^*) = \min_{\substack{V_p \subset V(G) \\ |V_p| = p}} \{F(V_p)\}.$$

The set V_p^* is called a *p-median* of G. We observe that in principle there is no restriction on where each facility may be located, that is, in a vertex or in a point along an edge. However, [30] showed that the set of vertices p-median always includes an absolute p-median, i.e., where some facilities could be located in points along the edges of G.

The portfolio optimization method we propose here is a clustering approach based on the p-median optimization problem to select portfolios by modeling the stock market as a network. In other words, we construct a portfolio by selecting each asset from the p clusters obtained by partitioning the (possibly complete) asset network $G = (V, E)$, where $V(G)$ is the set of stocks and the distances $d_{ij}, i, j \in V(G)$, are based on correlation coefficients ρ_{ij} between each pair of assets.

Following the seminal paper by Markowitz [40], the classical portfolio selection problem consists in determining the amount of capital to be invested in each asset of a given stock market. The problem is modeled as a bi-criteria mean-risk optimization problem as follows:

$$\max\{[\mu(\mathbf{x}), -\varrho(\mathbf{x})] | \mathbf{x} \in \Delta\},$$

where $\Delta = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{x} \geq 0, \sum_{i=1}^{n} x_i = 1\}$, x_i is the proportion of capital invested in asset i, $i = 1, \ldots, n$, $\mu(\mathbf{x})$ is the expected rate of return of a portfolio $\mathbf{x}$, and $\varrho : \mathbb{R}^n \to \mathbb{R}$ is a risk measure.

In this paper we show how the objective function of the classical p-median problem on networks can be used in order to filter the relevant information in a multivariate set of data. In particular, by means of an objective function of a classical location problem we are able to measure the clustering effect on the selected assets w.r.t. the non-selected ones. To this aim, given a network of assets $G = (V, E)$, we propose a tri-criteria optimization problem of the form:

$$\max\{[\mu(\mathbf{x}), -\varrho(\mathbf{x}), -F_p(\mathbf{x})] | \mathbf{x} \in \Delta,\ V_p \subset V(G), \text{ with } |V_p| = p\},$$

being $F_p(\mathbf{x}) = F(V_p)$. We note that we impose that a p-median of G must be a subset V_p of vertices (assets) of G and a vertex $j \in V_p$ can be considered as the *representative* asset of a given cluster. A vertex $i \notin V_p$ which is assigned to a vertex $j \in V_p$, is said to be represented by j. The aim is to take advantage of a p-median based clustering approach in order to diversify the portfolios. More precisely, on the one hand, with the p-median objective function, we invest in the representative (the median) of a cluster, thus avoiding to invest in more than one assets belonging to the same cluster (i.e., avoiding to invest in assets similar in terms of correlations). On the other hand, we are able to create diversified clusters and selecting one representative of each of them, that is, we are not selecting the more dissimilar assets but the representatives of dissimilar groups each one gathering similar assets. A detailed description of the behavior of our approach is presented in Section 4.2.

We provide a multi-objective mixed-integer mathematical programming formulation for dealing with

the above problem, that can be written as follows:

$$\max \quad \mu(\mathbf{x}) \tag{1a}$$
$$\min \quad \varrho(\mathbf{x}) \tag{1b}$$
$$\min \quad F_p(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} z_{ij} \tag{1c}$$
$$\text{s.t.} \quad \mathbf{x} \in \Delta \tag{1d}$$
$$\sum_{j=1}^{n} z_{jj} = p \tag{1e}$$
$$\sum_{j=1}^{n} z_{ij} = 1 \quad i = 1, \ldots, n \tag{1f}$$
$$z_{ij} \leq z_{jj} \quad i, j = 1, \ldots, n \tag{1g}$$
$$\ell_j z_{jj} \leq x_j \leq u_j z_{jj} \quad j = 1, \ldots, n \tag{1h}$$
$$z_{ij} \in \{0, 1\} \quad i, j = 1, \ldots, n \tag{1i}$$

where:

$$z_{jj} = \begin{cases} 1 & \text{if asset } j \text{ is selected in the portfolio and considered as a representative} \\ 0 & \text{otherwise} \end{cases}$$

$$z_{ij} = \begin{cases} 1 & \text{if asset } i \text{ is not selected in the portfolio but it is represented by asset } j \\ 0 & \text{otherwise.} \end{cases}$$

Constraint (1e) assures to select exactly p representatives (facilities); constraints (1f) guarantee that each asset belongs to exactly one cluster, whereas constraints (1g) assure that an asset i is represented by j only if j is selected as a representative of a cluster. Constraints (1h) tie together the location and portfolio selection problems. In fact, constraint (1h) states that if an asset j is selected as a representative, then we invest in asset x_j an amount of capital that must be at least equal to $\ell_j \geq 0$, to prevent holding of a position below a minimal allowable size, and at most $u_j \geq 0$, to prevent holding too large positions. In the following Γ denotes the feasible region determined by (1d)-(1i).

There exist different approaches for handling the multi-objective optimization problem (1a)-(1i). Most of the methods convert the given multi-objective program to a single-objective one after a scalarization of the different objective functions. Among the different approaches we consider the ϵ-constraint method. It consists in solving the problem with respect to one of its objective functions whereas the remaining objective functions are included as constraints parametrically varying their right-hand-sides with values [23]. Taking values ϱ_0 and μ_0 for the risk measure and the average return, respectively, our

problem becomes:

$$\min \quad \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} z_{ij} \tag{2a}$$

$$\text{s.t.} \quad (\mathbf{x}, \mathbf{z}) \in \Gamma \tag{2b}$$

$$\varrho(\mathbf{x}) \leq \varrho_0, \tag{2c}$$

$$\mu(\mathbf{x}) \geq \mu_0. \tag{2d}$$

This approach can also handle problems with integrality constraints by adequately choosing the different values of the right-hand-sides. From a portfolio selection viewpoint, as also shown in the seminal paper of Chang et al. [14], the presence of integrality constraints (e.g., cardinality constraints) can make the efficient frontier discontinuous, where the discontinuities might imply that there are certain returns which no rational investor would consider.

To analyze the complexity status of our location and portfolio selection problem, let us consider precisely the single-objective program (2a)-(2d). It clearly contains the classical p-median problem on networks as a special case. Assume $\varrho_0 = +\infty$ and $\mu_0 = -\infty$, $\ell_j = 0$ and $u_j = 1$, then constraints (2c) and (2d) can be ignored, i.e., a feasible portfolio can be always provided by setting $x_j = \frac{1}{p}$, for all j such that $z_{jj} = 1$, and problem (2a)-(2d), in fact, reduces to finding p facilities in order to minimize the sum of the distances between the opened facilities and the vertices of G. It is well-known that the p-median problem is NP-hard even in the case where the network is a planar graph of maximum vertex degree 3, all of whose edges and vertices have weight 1 [30]. Therefore, our location and portfolio selection problem is NP-hard, as well.

3 Solution method

In this section we present the solution technique used for obtaining solutions of the multi-objective optimization problem (1a)-(1i). We adopt the well-known ϵ-constraint method which is based on retaining only one of the three objectives and to turn all the others into constraints [23]. We retain the risk measure as objective, so that model (1a)-(1i) can be rewritten as follows:

$$\min \quad \varrho(\mathbf{x}) \tag{3a}$$

$$\text{s.t.} \quad \mu(\mathbf{x}) \geq \mu_0 \tag{3b}$$

$$F_p(\mathbf{x}) \leq F_p^0 \tag{3c}$$

$$(\mathbf{x}, \mathbf{z}) \in \Gamma \tag{3d}$$

where μ_0 and F_p^0 are, respectively, a lower bound for $\mu(\mathbf{x})$ and an upper bound for $F_p(\mathbf{x})$.

Problem (3a)-(3d) is a general *location and portfolio selection* problem in the sense that any risk measures can be considered as objective function. Nevertheless, in this paper we focus on some measures based only on the historical returns. In particular, we consider the Conditional Value at Risk (CVaR) as objective function [45]. We observe that other historical returns based measures could be used (Mean Absolute Deviation, Gini's mean difference [21], minimax objective function...). Our choice is motivated by the fact that this measure leads to an integer linear programming formulation of our problem instead of to nonlinear (e.g., quadratic) integer programs that are more difficult to solve. Moreover, CVaR is a well-establish risk measure in portfolio optimization (see, e.g., [36, 37] and the references therein).

In any case, it is not possible to completely ignore the correlation between the time series of returns of each asset that must be considered in order to quantify a measure of similarity between pairs of stocks. A widespread choice consists in quantifying the similarity between two assets with Pearson's correlation. Thus, the asset graph G represents the correlation structure between stocks where the weights/distances associated to each edge refer to the linear correlation between them [16]. Therefore, following a common practice in the literature about filtering procedures based on correlation clustering [49], we assign to each edge $(i,j) \in E(G)$ a similarity measure (distance) $d_{ij} = \sqrt{2(1-\rho_{ij})}$ where $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ is the correlation coefficient between assets' returns. We remark here that any alternative measures like the Kendall rank correlation coefficient or the Tail coefficient could be used as distances between assets (see, e.g., [16]). Hence, the correlation coefficients between stocks affect the location/clustering process of the model, while the CVaR measures the risk of the selected portfolios.

In the light of the above discussion, in the following we introduce problem (3a)-(3d) with the CVaR of a portfolio as objective function. First, we need to define the CVaR of a portfolio $\mathbf{x}$, and to this aim we follow the same notation introduced in [37]. Let us consider T different scenarios for the returns of the n assets. Each scenario t has associated a probability p_t, $t = 1, \ldots, T$, with $\sum_{t=1}^{T} p_t = 1$. Let r_{jt} be the rate of return of asset j at time t, $j = 1, \ldots, n$, $t = 1, \ldots, T$. Let μ_j be the average rate of return of asset j, i.e., $\mu_j = \sum_{t=1}^{T} p_t r_{jt}$, $j = 1, \ldots, n$. The rate of return at time t of a portfolio $\mathbf{x} = (x_1, \ldots, x_n) \in \Delta$ is

$$y_t(\mathbf{x}) = \sum_{j=1}^{n} r_{jt} x_j,$$

and the corresponding expected rate of return is

$$\mu(\mathbf{x}) = \sum_{j=1}^{n} \mu_j x_j.$$

The CVaR of a portfolio $\mathbf{x} = (x_1, \ldots, x_n) \in \Delta$ with tolerance level $\beta \in (0, 1]$ is defined as

$$M_\beta(\mathbf{x}) = \min_{u_t} \{ \frac{1}{\beta} \sum_{t=1}^{T} y_t(\mathbf{x}) u_t \mid \sum_{t=1}^{T} u_t = \beta, 0 \leq u_t \leq p_t \quad t = 1, \ldots, T \},$$

where $M_\beta(\mathbf{x})$ is the mean return of the portfolio taken over a given tolerance level $0 < \beta \leq 1$ of the worst scenarios probability. At optimality u_t is the percentage of the t-th worst return in $M_\beta(\mathbf{x})$. We observe that the above program is nonlinear, but we can overcome this difficulty by considering its dual with auxiliary linear equalities. Indeed, introducing a dual variable η that corresponds to the equation $\sum_{t=1}^{T} u_t = \beta$ and variables d_t^- corresponding to upper bounds on u_t, the dual, $M_\beta^D(\mathbf{x})$, of problem $M_\beta(\mathbf{x})$ is (for more details, the interested reader can refer to [37] pages 30-32)

$$M_\beta^D(\mathbf{x}) = \max_{\eta, d_t^-} \{ \eta - \frac{1}{\beta} \sum_{t=1}^{T} p_t d_t^- : d_t^- \geq \eta - y_t(\mathbf{x}), d_t^- \geq 0 \quad t = 1, \ldots, T, \eta \in \mathbb{R} \}.$$

We note that a similar approach based on the realizations of the portfolio rate of return was adopted in [6] to provide a Mixed-Integer Linear Programming formulation for the problem of finding an optimal mean/Value-at-Risk portfolio.

Finally, by substituting in (3a)-(3d) the objective (3a) with $M_\beta^D(\mathbf{x})$, our problem can be reformulated as the following Mixed-Integer Linear Programming (MILP) problem

$$\max \quad \eta - \frac{1}{\beta}\sum_{t=1}^{T} p_t d_t^- \tag{4a}$$

$$\text{s.t.} \quad d_t^- \geq \eta - y_t(\mathbf{x}) \quad t = 1, \ldots, T \tag{4b}$$

$$d_t^- \geq 0 \quad t = 1, \ldots, T \tag{4c}$$

$$\eta \in \mathbb{R} \tag{4d}$$

$$\mu(\mathbf{x}) \geq \mu_0 \tag{4e}$$

$$F_p(\mathbf{x}) \leq F_p^0 \tag{4f}$$

$$(\mathbf{x}, \mathbf{z}) \in \Gamma. \tag{4g}$$

Remark 1 *[25] In a p-median problem on networks, variable z_{ij}, $i \neq j$, takes value 1 in an optimal solution for a representative j such that $d_{ij} = \min_{k:z_{kk}=1}\{d_{ik}\}$.*

Remark 1 implies that in model (4a)-(4g) we can consistently reduce the number of binary variables from $O(n^2)$ to $O(n)$ by substituting (1i) in Γ with the following constraints

$$z_{jj} \in \{0, 1\} \quad j = 1, \ldots, n$$

$$z_{ij} \geq 0 \quad 1 \leq i, j \leq n, \; i \neq j.$$

3.1 Selection of F_p^0

The selection of the lower bound on the portfolio expected return μ_0 in model (4a)-(4g) is, evidently, of the taste of the investor. However, how to control the effect of clusterization by means of the upper bound F_p^0 is not obvious. In this section we describe how to compute F_p^0 in (4a)-(4g).

Consider the problem

$$\min \quad F_p(\mathbf{x}) \tag{5a}$$

$$\text{s.t.} \quad \mu(\mathbf{x}) \geq \mu_0 \tag{5b}$$

$$(\mathbf{x}, \mathbf{z}) \in \Gamma \tag{5c}$$

and let F_p^ℓ be its optimal value. Note that F_p^ℓ is the minimum value for F_p^0 that makes problem (4a)-(4g) feasible.

Let $\mathcal{S}$ be the set of feasible solutions defined by (1d) together with (4b)-(4d). Consider now the problem

$$\max \quad \eta - \frac{1}{\beta}\sum_{t=1}^{T} p_t d_t^- \tag{6a}$$

$$\text{s.t.} \quad (\mathbf{x}, \mathbf{d}^-, \eta) \in \mathcal{S} \tag{6b}$$

$$\mu(\mathbf{x}) \geq \mu_0 \tag{6c}$$

$$\sum_{j=1}^{n} z_j = p \tag{6d}$$

$$\ell_j z_j \leq x_j \leq u_j z_j \qquad j = 1, \ldots, n \tag{6e}$$

$$z_j \in \{0, 1\} \qquad j = 1, \ldots, n \tag{6f}$$

and let X_p be the set of the p selected assets in an optimal solution. Then, $F_p(X_p) = F_p^u$ is the tightest upper bound for F_p^0, that is, for all $F_p^0 \geq F_p^u$ the objective $F_p(\mathbf{x})$ is negligible in (4a)-(4g).

Finally, given $\lambda \in [0, 1]$, the parameter F_p^0 can be computed as follows

$$F_p^0 = \lambda F_p^\ell + (1 - \lambda) F_p^u.$$

In this way, the selection of the parameter λ allows us to control the required clustering effect, from the highest one, with $\lambda = 1$, to the lowest one, with $\lambda = 0$ (see Section 4.2 for more details).

4 Experimental results

This section presents an empirical analysis with the aim of evaluating the performance of the portfolios selected by our location and selection model (4a)-(4g). In addition, we also compare our model with the pure CVaR approach [45], regarded as the benchmark program, and the CVaR model with cardinality constraints.

The pure CVaR model is [36, 45]

$$\max \quad \eta - \frac{1}{\beta}\sum_{t=1}^{T} p_t d_t^- \tag{7a}$$

$$\text{s.t.} \quad (\mathbf{x}, \mathbf{d}^-, \eta) \in \mathcal{S} \tag{7b}$$

$$\mu(\mathbf{x}) \geq \mu_0 \tag{7c}$$

and the CVaR model with a limited number of assets (CVaR-CC) is precisely the above model (6a)-(6f) (see, e.g., [13]).

4.1 Data sets

We test all the above portfolio selection strategies on some real-world datasets belonging to the major stock markets across the world. We consider the following datasets:

1. DJIA (Dow Jones Industrial Average, USA), containing 28 assets and 1353 price observations (period: 07/05/1990 - 04/04/2016);

2. EUROSTOXX50 (Europe's leading blue-chip index, EU), containing 49 assets and 729 price observations (period: 22/04/2002-04/04/2016);

3. FTSE100 (Financial Times Stock Exchange, UK), containing 83 assets and 625 price observations (period: 19/04/2004-04/04/2016);

4. SP500 (Standard & Poor's, USA), containing 442 assets and 573 observations (period: 18/04/2005-04/04/2016).

Each dataset consists of weekly prices data. To evaluate the performance of our models in practice, we divide the observations in two sets, where the first one is regarded as the past (in-sample window), and so it is known, and the rest is regarded as the future (out-of-sample window), supposed unknown at the time of portfolio selection. The in-sample window is used for selecting the portfolio, while the out-of-sample one is used for testing the performance of the selected portfolio. In particular, in our experiments we use a *rolling time window* scheme allowing for the possibility of rebalancing the portfolio composition during the holding period, at fixed intervals. Following [28, 36], for each dataset we adopt a period of 104 weeks (two years) as in-sample window and, in a first experimental set-up, we consider 52 weeks (one year) as out-of-sample, with rebalancing allowed every 52 weeks. In this phase, for each value of p, each model solves overall 55 problems (24 for DJIA, 12 for EUROSTOXX50, 10 for FTSE100 and 9 for SP500).

Let $0, 1, \ldots, T$ be the observations in the in-sample window; for each dataset, to compute the $T \times n$ matrix of the historical returns, we consider the time series of the prices of each of the n stocks and denote by $P_i(t)$ the price of the ith asset at time t, $t = 0, \ldots, T$. The ith asset return at time t is computed as $r_{it} = \frac{P_i(t)-P_i(t-1)}{P_i(t-1)}$, with $t = 1, \ldots, T$. The value r_{it} measures the change in the value of a specific asset i at time t. Since in model (4a)-(4g), for a fixed time t, we refer to the rate of return of a portfolio $\mathbf{x}$, namely $y_t(\mathbf{x})$, r_{it} is the correct measure for determining the return of a portfolio computed as the weighted sum of the returns of the individual securities. On the other hand, as is standard practice in the literature, to capture the degree of correlation between pairs of stocks in a market over time, it seems more appropriate to compute, for each asset i, the rate of return with a continuous capitalization. This leads to referring to the series of the in-sample logarithmic returns of each asset i computed as $R_{it} = \ln P_i(t) - \ln P_i(t-1)$, with $t = 1, \ldots, T$ (see, e.g., [38, 44]). We note that with relatively small time intervals (daily, weekly), the logarithmic return R_{it} is approximately equal to the (discrete) return r_{it} over the same interval, so that the two values can be considered roughly similar.

Finally, we compute the distances d_{ij} between assets i and j, $i, j = 1, \ldots, n$, as $d_{ij} = \sqrt{2(1-\rho_{ij})}$ where ρ_{ij} is the Pearson correlation coefficient between assets' logarithmic returns. We do not set any threshold for the correlation coefficients, thus the resulting asset graph G is a complete weighted graph.

As in several other papers (see, e.g., [11] and the references therein) the values of p for the number of representatives, that are also used for bounding the number of assets in a portfolio, are set to $p = 5, 10, 15, 20$. We choose $\ell_i = \frac{1}{n}$ and $u_i = 1$ as lower and upper bounds for the minimum and maximum allowable amount of capital invested in each asset i, $i = 1, \ldots, n$. We decided to choose $\ell_i > 0$, $i = 1, \ldots, n$, in order to invest a positive amount of money in each of the p assets selected by our MILP model. In this way, we guarantee that we are fairly comparing a portfolio formed by exactly p assets provided by our MILP program with the ones found by the CVaR-CC model.

In portfolio selection problems, the out-of-sample performance of a portfolio is generally evaluated by using some performance measures. In our experiments we consider the following measures (see, e.g., [9])

1. *Sharpe Ratio* (Sh) ([46, 47]): it is defined as the ratio between the average of the out-of-sample rates of return of a portfolio $\mathbf{x}$, that we denote by $\mu^{\text{out}}(\mathbf{x})$, minus a constant risk free rate of return r_f (that we set equal to 0), and its standard deviation, namely:

$$\frac{E[\mu^{\text{out}}(\mathbf{x}) - r_f]}{\sigma(\mu^{\text{out}}(\mathbf{x}))}.$$

   The larger is the value of the index, the better is the portfolio performance.

2. *Average return* (Av): it is defined as the average $E[\mu^{\text{out}}(\mathbf{x})]$ of the out-of-sample returns of a portfolio.

The models have been implemented in MATLAB R2017A and they make calls to XPRESS solver version 8.5 for solving the MILP programs. All experiments were run in a computer DellT5500 with a processor Intel(R) Xeon(R) with a CPU X5690 at 3.75 GHz and 48 GB of RAM memory.

We report the out-of-sample performances in the following tables. In each table, the first column refers to the models implemented for the comparisons. In particular, with $\lambda = 0.1, \ldots, 1$ we refer to our MILP model (4a)-(4g) for the different values of λ used for finding the parameter F_p^0 which bounds the function $F_p(\mathbf{x})$ in (4a)-(4g) (see Section 3.1). CVaR-CC and CVaR refer to the cardinality constraint CVaR model (6a)-(6f) and to the pure CVaR model (7a)-(7c), respectively. We observe that the CVaR-CC model corresponds, in fact, to our MILP model when $\lambda = 0$. In all the models, we choose the average index return in each in-sample period as the lower bound μ_0 on the expected rate of return. We also consider the out-of-sample performance of the market index (Index) as benchmark. For each value of $p = 5, 10, 15, 20$, in the columns we report the out-of-sample average return (Av-p) and the value of the Sharpe Ratio (Sh-p). In addition, for the first three datasets (DJIA, EUROSTOXX50 and FTSE100) we also report the average in-sample solution times. Finally, for the pure CVaR model, we also report the average number of assets in the optimal in-sample portfolios.
For the SP500 dataset, due to the size of the corresponding MILP model, we were unable to find an optimal solution in reasonable times. Thus, we set a time limit of 7200 seconds, and in the corresponding table we report the average percentage GAP (GAP%) in place of the solution times. Furthermore, we adopt a special methodology for solving the model in each in-sample period. Note that MILP model (4a)-(4g) for $\lambda = 1$ can be solved in two easy steps: first, solve problem (5a)-(5c) and then, after having fixed the binary variables z_j according to the p-median solution found, solve problem (6a)-(6c). Note also that the optimal solution of problem (4a)-(4g) for $\lambda = 1$ is a feasible solution of problem (4a)-(4g) for $\lambda = 0.9$, thus it can be used as initial solution in the branching procedure in order to get better solutions in smaller times. The same applies for a feasible solution of problem (4a)-(4g) for $\lambda = 0.9$ and problem (4a)-(4g) for $\lambda = 0.8$, and so on.

In the tables, in bold we provide the best values of the out-of-sample performance measures for each value of p. In order to highlight the effectiveness of our MILP approach, for each p and value of λ, we write the values of the two performance indexes in *italic* if they are both better than the corresponding values of the other models (CVaR-CC, CVaR, Index). In this way we can have a count of how many times our approach outperforms the other competing portfolio selection models.
From Tables 1-4, except for two cases (EUROSTOXX50 with 5 assets and SP500 with 10 assets) there always exists at least one value of λ for which our MILP model outperforms the other competing models, i.e., CVaR-CC, CVaR and the market index. We also note that, as expected, for the SP500 dataset, the GAP values, which are quite high for $p = 5, 10$, tends to become very small when p increases. In fact, for $p = 15, 20$ our MILP model produces in-sample solutions that we suppose are very close to the

Table 1: Average returns and Sharpe ratios for the DJIA dataset: in-sample 104 weeks; out-of-sample 52 weeks with a rebalancing allowed every 52 weeks.

Model	Av-5 ($\cdot 10^{-3}$)	Sh-5	Time-5	Av-10 ($\cdot 10^{-3}$)	Sh-10	Time-10	Av-15 ($\cdot 10^{-3}$)	Sh-15	Time-15	Av-20 ($\cdot 10^{-3}$)	Sh-20	Time-20
CVaR-CC	2.279	0.104	0.290	1.796	0.085	0.117	1.874	0.089	0.465	2.112	0.098	0.425
λ=0.1	2.064	0.091	0.587	1.848	0.087	0.620	*1.903*	*0.089*	0.566	*2.162*	*0.101*	0.907
λ=0.2	1.927	0.083	0.536	*1.882*	*0.089*	0.628	*1.997*	*0.094*	1.144	*2.178*	*0.102*	0.894
λ=0.3	2.088	0.091	0.718	1.820	0.086	0.597	*1.892*	*0.090*	0.834	*2.163*	*0.101*	1.540
λ=0.4	2.206	0.097	0.693	1.770	0.084	0.986	*2.034*	*0.096*	1.061	*2.225*	*0.104*	1.173
λ=0.5	2.333	0.103	0.642	1.820	0.084	0.779	*2.001*	*0.094*	0.722	*2.235*	*0.104*	1.306
λ=0.6	*2.338*	*0.104*	0.630	*1.913*	*0.088*	0.741	*1.963*	*0.092*	1.083	*2.166*	*0.101*	1.179
λ=0.7	***2.366***	***0.105***	0.685	*1.985*	*0.091*	0.731	***2.100***	***0.099***	0.949	*2.122*	*0.099*	0.981
λ=0.8	1.933	0.086	0.654	*1.989*	*0.091*	0.623	*2.068*	*0.097*	1.044	*2.276*	*0.106*	1.177
λ=0.9	2.285	0.099	0.667	*2.006*	*0.091*	0.680	*2.060*	*0.097*	0.945	*2.157*	*0.100*	1.670
λ=1	2.179	0.088	0.082	***2.321***	***0.105***	0.130	*2.098*	*0.097*	0.191	***2.379***	***0.108***	0.139

	Av ($\cdot 10^{-3}$)	Sh	Time	n. of assets
CVaR	1.856	0.087	0.003	8.6
Index	1.608	0.068	-	28

Table 2: Average returns and Sharpe ratios for the EUROSTOXX50 dataset: in-sample 104 weeks, out-of-sample 52 weeks with a rebalancing allowed every 52 weeks.

Model	Av-5 ($\cdot 10^{-3}$)	Sh-5	Time-5	Av-10 ($\cdot 10^{-3}$)	Sh-10	Time-10	Av-15 ($\cdot 10^{-3}$)	Sh-15	Time-15	Av-20 ($\cdot 10^{-3}$)	Sh-20	Time-20
CVaR-CC	***2.103***	***0.078***	0.327	1.955	0.073	0.077	1.939	0.075	0.102	2.008	0.077	0.150
λ=0.1	2.008	0.075	1.239	1.946	0.074	1.528	1.908	0.074	1.383	*2.020*	*0.077*	2.723
λ=0.2	2.051	0.076	1.108	*1.977*	*0.075*	2.134	1.936	0.075	1.888	1.973	0.076	2.203
λ=0.3	1.979	0.073	1.361	*1.974*	*0.074*	2.336	*1.971*	*0.076*	3.167	1.993	0.076	2.720
λ=0.4	1.972	0.073	1.670	1.950	0.074	1.596	1.925	0.074	1.579	1.981	0.075	2.915
λ=0.5	1.771	0.063	1.845	*2.047*	*0.077*	1.610	*2.037*	***0.078***	2.699	*2.039*	*0.078*	3.446
λ=0.6	1.874	0.069	1.977	1.910	0.072	1.582	1.944	0.074	2.512	*2.054*	*0.079*	3.067
λ=0.7	1.990	0.073	2.137	*2.005*	*0.076*	2.145	*2.006*	*0.076*	3.593	*2.051*	*0.078*	4.445
λ=0.8	1.766	0.064	2.151	***2.090***	***0.078***	2.810	***2.059***	***0.078***	2.770	*2.063*	*0.078*	3.749
λ=0.9	1.714	0.062	2.068	*1.979*	*0.074*	2.134	*2.045*	*0.076*	3.350	***2.111***	***0.080***	5.329
λ=1	0.868	0.030	0.274	1.967	0.065	0.241	1.803	0.067	0.153	1.887	0.070	0.285

	Av ($\cdot 10^{-3}$)	Sh	Time	n. of assets
CVaR	1.922	0.072	0.002	7.2
Index	0.509	0.017	-	49

Table 3: Average returns and Sharpe ratios for the FTSE100 dataset: in-sample 104 weeks, out-of-sample 52 weeks with a rebalancing allowed every 52 weeks.

Model	Av-5 ($\cdot 10^{-3}$)	Sh-5	Time-5	Av-10 ($\cdot 10^{-3}$)	Sh-10	Time-10	Av-15 ($\cdot 10^{-3}$)	Sh-15	Time-15	Av-20 ($\cdot 10^{-3}$)	Sh-20	Time-20
CVaR-CC	2.045	0.091	0.442	1.492	0.069	0.317	1.623	0.076	0.329	1.696	0.080	0.225
λ=0.1	2.066	0.089	17.495	*1.581*	*0.074*	6.434	1.542	0.072	2.732	*1.698*	*0.080*	2.309
λ=0.2	1.830	0.079	20.030	*1.687*	*0.078*	8.577	1.593	0.075	2.975	1.686	0.079	2.546
λ=0.3	1.627	0.071	24.037	*1.596*	*0.075*	24.238	1.540	0.072	2.658	1.653	0.078	3.724
λ=0.4	1.807	0.079	29.242	*1.623*	*0.075*	14.249	*1.639*	*0.077*	2.809	1.597	0.076	2.605
λ=0.5	***2.104***	***0.092***	27.456	*1.683*	*0.079*	17.795	*1.760*	*0.082*	4.379	1.635	0.077	3.760
λ=0.6	1.921	0.083	26.529	***1.870***	***0.085***	16.946	*1.698*	*0.079*	6.225	*1.699*	*0.080*	3.393
λ=0.7	1.628	0.066	22.059	*1.681*	*0.077*	9.477	1.590	0.074	6.440	*1.710*	*0.081*	2.966
λ=0.8	1.310	0.053	19.084	*1.808*	*0.081*	13.817	***1.870***	***0.087***	5.646	*1.817*	*0.085*	3.238
λ=0.9	1.499	0.059	10.887	*1.716*	*0.077*	8.265	*1.785*	*0.082*	5.811	*1.845*	*0.086*	3.873
λ=1	0.925	0.026	2.271	1.480	0.057	0.873	1.754	0.069	1.261	***2.113***	***0.095***	0.791

	Av ($\cdot 10^{-3}$)	Sh	Time	n. of assets
CVaR	1.528	0.071	0.008	11.7
Index	0.040	0.015	-	83

Table 4: Average returns and Sharpe ratios for the SP500 dataset: in-sample 104 weeks, out-of-sample 52 weeks with a rebalancing allowed every 52 weeks.

Model	Av-5 ($\cdot 10^{-3}$)	Sh-5	GAP%-5	Av-10 ($\cdot 10^{-3}$)	Sh-10	GAP%-10	Av-15 ($\cdot 10^{-3}$)	Sh-15	GAP%-15	Av-20 ($\cdot 10^{-3}$)	Sh-20	GAP%-20
CVaR-CC	2.224	0.094	0.059	***2.515***	***0.108***	0.059	2.187	0.095	0.044	2.130	0.093	0.007
λ=0.1	1.970	0.080	23.634	2.452	0.103	4.371	***2.266***	***0.098***	0.302	2.165	0.095	0.058
λ=0.2	1.877	0.075	24.092	2.466	0.104	4.814	*2.255*	***0.098***	0.304	*2.170*	*0.095*	0.063
λ=0.3	2.028	0.082	25.405	2.478	0.104	5.243	2.186	0.094	0.404	*2.170*	*0.095*	0.065
λ=0.4	1.881	0.075	31.787	2.402	0.103	6.853	2.174	0.094	0.540	*2.181*	*0.095*	0.058
λ=0.5	1.861	0.073	34.658	2.271	0.097	8.286	2.182	0.095	0.629	*2.193*	*0.096*	0.037
λ=0.6	1.980	0.077	35.634	2.024	0.083	9.302	2.077	0.089	1.293	*2.183*	*0.095*	0.097
λ=0.7	***2.386***	***0.098***	43.565	2.182	0.089	14.402	2.109	0.090	2.527	*2.216*	*0.096*	0.088
λ=0.8	2.052	0.084	48.679	1.796	0.076	22.566	2.073	0.087	4.250	2.116	0.092	0.463
λ=0.9	*2.239*	***0.098***	54.367	1.906	0.076	27.719	2.113	0.088	11.997	***2.284***	***0.098***	2.948
λ=1	2.010	0.067	0	1.364	0.046	0	1.784	0.067	0	1.291	0.056	0

	Av ($\cdot 10^{-3}$)	Sh	Time	n. of assets
CVaR	2.168	0.094	0.033	15.3
Index	1.137	0.040	-	442

Journal Pre-proof

optimal ones, but XPRESS solver was unable to certify their optimality. This would suggest to develop an effective heuristic procedure that combines the good features of location and portfolio selection models in order to obtain optimal or near-optimal in-sample portfolios that will likely produce better out-of-sample performances, in particular, for large size financial datasets.

In order to highlight the effectiveness of our approach, we also decided to provide a second set of experiments. In this case, we adopt a period of 104 weeks as in-sample window and 12 weeks (three months) as out-of-sample, with rebalancing allowed every 12 weeks (see also [36]). For each value of p, each model solves overall 204 problems (104 for DJIA, 55 for EUROSTOXX50, 45 for FTSE100). Note that, in this case, we did not consider the SP500 dataset, since we are now interested in comparing our approach with the other programs when our MILP model is able to find optimal (certified) in-sample solutions. In any case, we observed that some preliminary results with the SP500 dataset are in line with the ones reported in Table 4.

Following the results in Tables 5-7, it is clear that, when the MILP model is able to find optimal in-sample solutions, our approach proves to be efficient both from a computational times of view, and always dominates the other methods. To conclude this section, we note that, quite surprisingly, in all the experimental analysis the pure CVaR model is always dominated by the other methods.

Table 5: Average returns and Sharpe ratios for the DJIA dataset: in-sample 104 weeks, out-of-sample 12 weeks with a rebalancing allowed every 12 weeks

Model	Av-5 ($\cdot 10^{-3}$)	Sh-5	Time-5	Av-10 ($\cdot 10^{-3}$)	Sh-10	Time-10	Av-15 ($\cdot 10^{-3}$)	Sh-15	Time-15	Av-20 ($\cdot 10^{-3}$)	Sh-20	Time-20
CVaR-CC	1.530	0.070	0.276	1.499	0.071	0.137	1.752	0.083	0.300	2.088	0.098	0.371
λ=0.1	*1.726*	*0.077*	0.596	*1.608*	*0.077*	0.712	1.746	0.084	0.952	*2.198*	*0.104*	0.896
λ=0.2	*1.851*	*0.083*	0.540	*1.616*	*0.077*	0.681	*1.766*	*0.085*	0.940	*2.145*	*0.100*	0.937
λ=0.3	*1.823*	*0.081*	0.630	1.587	0.077	0.726	*1.805*	*0.086*	1.244	*2.173*	*0.102*	1.233
λ=0.4	*1.756*	*0.078*	0.630	1.511	0.072	0.701	*1.786*	*0.085*	1.076	*2.199*	*0.102*	0.875
λ=0.5	*1.691*	*0.074*	0.686	1.529	0.072	0.798	*1.867*	*0.088*	0.989	*2.200*	*0.103*	1.399
λ=0.6	*1.765*	*0.077*	0.697	*1.735*	*0.082*	0.857	*1.879*	*0.090*	0.763	*2.239*	*0.104*	1.233
λ=0.7	1.595	0.070	0.666	*1.664*	*0.079*	0.768	*1.869*	*0.088*	0.942	*2.193*	*0.102*	1.495
λ=0.8	*1.625*	*0.071*	0.710	*1.629*	*0.077*	0.791	*1.987*	*0.094*	1.066	***2.356***	***0.110***	1.189
λ=0.9	*1.855*	*0.079*	0.683	*1.644*	*0.076*	0.828	*1.877*	*0.089*	1.089	*2.135*	*0.099*	1.302
λ=1	***2.103***	***0.088***	0.101	***2.127***	***0.098***	0.154	***2.158***	***0.101***	0.203	*2.236*	*0.103*	0.118

	Av ($\cdot 10^{-3}$)	Sh	Time	n. of assets
CVaR	1.518	0.071	0.003	8.7
Index	1.608	0.068	-	28

Table 6: Average returns and Sharpe ratios for the EUROSTOXX50 dataset: in-sample 104 weeks, out-of-sample 12 weeks with a rebalancing allowed every 12 weeks

Model	Av-5 ($\cdot 10^{-3}$)	Sh-5	Time-5	Av-10 ($\cdot 10^{-3}$)	Sh-10	Time-10	Av-15 ($\cdot 10^{-3}$)	Sh-15	Time-15	Av-20 ($\cdot 10^{-3}$)	Sh-20	Time-20
CVaR-CC	1.863	0.076	0.162	1.796	0.074	0.104	1.804	0.075	0.148	1.964	0.082	0.202
λ=0.1	*1.975*	*0.081*	1.314	*1.816*	*0.075*	1.418	*1.851*	*0.077*	2.031	1.928	0.080	2.022
λ=0.2	*2.004*	*0.084*	1.490	*1.808*	*0.075*	1.297	*1.844*	*0.077*	1.529	*1.971*	*0.082*	2.510
λ=0.3	*2.102*	*0.086*	1.675	*1.821*	*0.075*	1.712	*1.917*	*0.080*	2.247	1.941	0.080	2.066
λ=0.4	*2.117*	*0.087*	1.652	*1.760*	*0.073*	1.655	*1.899*	*0.079*	2.105	1.929	0.080	2.654
λ=0.5	*2.224*	***0.092***	1.736	*1.779*	*0.073*	1.601	*1.887*	*0.078*	2.295	*2.045*	*0.085*	3.068
λ=0.6	*2.061*	*0.085*	1.930	*1.866*	*0.077*	1.792	*1.896*	*0.078*	2.758	***2.074***	***0.086***	2.693
λ=0.7	*1.921*	*0.080*	2.065	1.794	0.074	2.407	*1.951*	*0.080*	3.260	*2.060*	*0.085*	3.665
λ=0.8	*2.069*	*0.084*	1.930	*1.931*	*0.079*	2.414	*2.021*	*0.083*	3.295	*2.061*	*0.085*	4.081
λ=0.9	1.826	0.072	1.933	*1.983*	*0.082*	2.320	***2.106***	***0.086***	4.195	*2.061*	*0.084*	4.184
λ=1	***2.228***	*0.077*	0.368	***2.207***	***0.090***	0.272	*1.993*	*0.079*	0.211	*2.069*	*0.083*	0.216

	Av ($\cdot 10^{-3}$)	Sh	Time	n. of assets
CVaR	1.745	0.071	0.003	6.9
Index	0.712	0.024	-	49

Table 7: Average returns and Sharpe ratios for the FTSE100 dataset: in-sample 104 weeks, out-of-sample 12 weeks with a rebalancing allowed every 12 weeks

Model	Av-5 ($\cdot 10^{-3}$)	Sh-5	Time-5	Av-10 ($\cdot 10^{-3}$)	Sh-10	Time-10	Av-15 ($\cdot 10^{-3}$)	Sh-15	Time-15	Av-20 ($\cdot 10^{-3}$)	Sh-20	Time-20
CVaR-CC	2.417	0.113	0.446	2.235	0.106	0.337	2.273	0.109	0.273	2.270	0.111	0.316
λ=0.1	*2.771*	*0.123*	13.531	2.180	0.103	4.895	2.275	0.109	2.637	2.207	0.108	3.033
λ=0.2	*2.670*	*0.119*	19.999	*2.322*	***0.110***	5.695	*2.325*	***0.111***	2.486	2.257	0.109	3.287
λ=0.3	2.290	0.099	20.967	2.138	0.101	7.448	2.236	0.107	2.898	2.198	0.107	3.274
λ=0.4	2.197	0.097	24.230	2.125	0.100	8.006	2.195	0.104	2.864	2.183	0.105	2.909
λ=0.5	2.495	0.112	24.582	2.047	0.096	7.776	2.177	0.103	3.171	2.120	0.102	3.062
λ=0.6	2.013	0.091	26.766	2.131	0.100	8.964	2.175	0.103	3.318	2.151	0.104	3.257
λ=0.7	*2.626*	*0.117*	24.267	2.190	0.103	9.490	2.068	0.098	3.841	2.170	0.104	2.976
λ=0.8	***3.025***	***0.131***	17.133	2.283	0.107	9.134	2.128	0.101	4.018	2.187	0.104	3.770
λ=0.9	1.459	0.059	10.543	**2.341**	0.107	8.135	2.109	0.099	4.700	2.221	0.104	3.893
λ=1	0.876	0.029	1.803	1.529	0.065	2.025	**2.452**	0.107	1.311	***2.560***	***0.118***	1.099

	Av ($\cdot 10^{-3}$)	Sh	Time	n. of assets
CVaR	2.315	0.109	0.009	11
Index	0.567	0.021	-	83

4.2 Out-of-sample performance evaluation

To better understand the effectiveness of our approach and how it is related to the clusterization technique, in this section we provide a more detailed analysis about the out-of-sample behavior of our MILP model w.r.t. the competing ones. In particular, we focus our analysis on the comparison between the MILP model and the CVaR-CC one since, as shown in the previous tables, the pure CVaR model is always dominated by the other approaches.

As reported in Section 2, the aim of this paper is to take advantage of a p-median based clustering approach in order to obtain portfolios that avoid to invest in more than one asset belonging to the same cluster (i.e., avoiding to invest in assets similar in terms of correlations). Thus, by selecting one representative of each cluster, we are, in fact, investing in the representatives of dissimilar groups of assets while minimizing a given risk measure. This clusterization and selection strategy should produce better portfolios, for example, in terms of average returns and portfolio values. In particular, for the out-of-sample returns this is confirmed by observing that in Tables 1-7, the best values of the average return is consistently obtained for values of lambda greater than 0.5. The only exception is Table 4 that refers to the SP500 dataset, where for $p = 10$ our MILP model is dominated by the CVaR-CC one and for $p = 15$ the best average return is for $\lambda = 0.1$. This can be due to the fact that for this dataset we were not able to find the optimal in-sample portfolio, in particular for values of λ greater than 0.5 for which we observe that the greater the value of λ the greater the corresponding value of the average percentage GAP. In any case, the parameter λ seems to have a meaningful effect on the out-of-sample performance of our portfolios. In fact, recall that parameter λ allows us to control the required clustering effect, from the highest one ($\lambda = 1$) to the lowest one ($\lambda = 0$) (see Subection 3.1). Actually, when $\lambda = 0$ we have $F_p(X_p) = F_p^u$ which is the tightest upper bound for F_p^0, that is, for all $F_p^0 \geq F_p^u$ the objective $F_p(\mathbf{x})$ is negligible in our model, and we are, in fact, solving the CVaR-CC model which is contained in our MILP program as a special case (i.e., when $F_p^0 \geq F_p^u$). When $\lambda = 1$, we have $F_p(X_p) = F_p^\ell$ and the p-median objective function is of greater importance.

For the sake of brevity and because the conclusions are similar in the remaining datasets, let us just consider the two datasets, EUROSTOXX50 and SP500 with a rebalancing allowed every 52 weeks, for which in two cases (see Tables 2 and 4 for $p = 5$ and $p = 10$, respectively) CVaR-CC model performs better than our MILP model. In the following figures, for each of these two datasets we compare the weekly out-of-sample values of the portfolios computed by our MILP model and the CVaR-CC model in the cases $p = 5$ and $p = 10$ for EUROSTOXX50 and $p = 10$ and $p = 20$ for SP500. We consider the cases $p = 10$ and $p = 20$ for EUROSTOXX50 and SP500, respectively, in order to show the differences in the out-of-sample performance when a portfolio provided by our MILP model dominates the one produced by the CVaR-CC model.

We note that even if in Tables 2 and 4 CVaR-CC provides portfolios with better average returns than our MILP model, we observe that, in this case, the two graphics (see Figure 1(a) and Figure 1(c)) are very close to each other, meaning that the weekly out-of-sample performances are, in fact, very similar with relatively small differences over time. In the two remaining cases (see Figure 1(b) and Figure 1(d)), our portfolios more clearly outperform those produced by the CVaR-CC model, in particular, when the clustering effect is more evident ($\lambda = 0.8$ and $\lambda = 0.9$, respectively). A similar behavior of the portfolios provided by our model is observed for all the other datasets (figures for all the other datasets are available upon requests).

To better analyze the clusterization effect, we also present two tables where, for all the datasets with a rebalancing allowed every 52 weeks, we report the average of the sum of the absolute differences (over the out-of-sample periods) between a portfolio $\mathbf{x}$ provided by our MILP model and a portfolio $\mathbf{x}'$ found by the CVaR-CC model. This index is computed as $\frac{1}{2}\sum_{i=1}^{n}|x_i - x_i'|$, and measures the differences in the asset weights composition between the two portfolios $\mathbf{x}$ and $\mathbf{x}'$. The higher the index value the more different are the two portfolios. The two following tables report separately the results for the DJIA and EUROSTOXX50 datasets (small size problems) and the results for the FTSE100 and SP500 datasets (medium and large size problems).

Model	**DJIA**				**EUROSTOXX50**			
	$p = 5$	$p = 10$	$p = 15$	$p = 20$	$p = 5$	$p = 10$	$p = 15$	$p = 20$
λ=0.1	0.20	0.14	0.10	0.08	0.13	0.08	0.06	0.05
λ=0.2	0.25	0.15	0.11	0.09	0.14	0.09	0.06	0.06
λ=0.3	0.20	0.15	0.11	0.08	0.12	0.08	0.07	0.07
λ=0.4	0.21	0.17	0.12	0.11	0.16	0.09	0.08	0.08
λ=0.5	0.21	0.19	0.13	0.12	0.16	0.10	0.08	0.08
λ=0.6	0.23	0.23	0.17	0.12	0.15	0.12	0.10	0.09
λ=0.7	**0.26**	0.22	**0.19**	0.14	0.15	0.13	0.10	0.10
λ=0.8	0.42	0.28	0.23	0.16	0.24	**0.13**	**0.12**	0.13
λ=0.9	0.56	0.34	0.28	0.19	0.31	0.18	0.16	**0.17**
λ=1	0.75	**0.56**	0.37	**0.24**	0.84	0.55	0.33	0.30

Table 8: Average of the sum of the absolute differences over the out-of-sample periods for the DJIA and EUROSTOXX50 datasets. In-sample 104 weeks, out-of-sample 52 weeks with a rebalancing allowed every 52 weeks. Boldfaced entries in the table point out the best portfolios found as previously shown in Tables 1-2.

The values in the tables explain the effect of the change of the λ parameter in the portfolios composition w.r.t. the corresponding ones produced by the CVaR-CC model for all values of $p = 5, 10, 15, 20$. For instance, for DJIA with $p = 10$ and $\lambda = 1$, the portfolio has changed in a 56% of its composition w.r.t. the portfolio provided by the CVaR-CC model for $p = 10$. As shown in the above tables, when we increase λ, thus imposing a higher clusterization effect, we observe a sort of monotonicity in the way in which the composition of our portfolios change. In fact, in those cases where our model obtains its best performance for a low value of λ (e.g., for SP500 with p = 15), is because the portfolios' composition is similar to the one provided by the CVaR-CC model. Recall that for λ approaching to 0 we are actually solving the CVaR-CC model which is contained in our MILP program as a special case. On the other hand, when the best performance is achieved with a high value of λ (typically for $\lambda \geq 0.5$), we observe that our portfolios have changed considerably in their composition w.r.t. the ones provided by

Model	FTSE100				SP500			
	$p=5$	$p=10$	$p=15$	$p=20$	$p=5$	$p=10$	$p=15$	$p=20$
λ=0.1	0.19	0.10	0.07	0.05	0.41	0.12	**0.08**	0.02
λ=0.2	0.23	0.12	0.06	0.05	0.43	0.14	0.08	0.02
λ=0.3	0.22	0.14	0.07	0.06	0.45	0.17	0.08	0.02
λ=0.4	0.17	0.13	0.10	0.07	0.49	0.26	0.10	0.04
λ=0.5	**0.22**	0.13	0.12	0.11	0.48	0.24	0.10	0.03
λ=0.6	0.29	**0.17**	0.10	0.09	0.49	0.26	0.12	0.06
λ=0.7	0.37	0.18	0.15	0.12	**0.59**	0.31	0.17	0.06
λ=0.8	0.49	0.22	**0.19**	0.16	0.58	0.38	0.19	0.09
λ=0.9	0.72	0.34	0.21	0.19	0.72	0.48	0.29	**0.22**
λ=1	1.00	0.96	0.81	**0.63**	1.00	1.00	1.00	1.00

Table 9: Average of the sum of the absolute differences over the out-of-sample periods for the FTSE100 and SP500 datasets. In-sample 104 weeks, out-of-sample 52 weeks with a rebalancing allowed every 52 weeks. Boldfaced entries in the table point out the best portfolios found as previously shown in Tables 3-4.

the CVaR-CC model which are also dominated by ours (see, e.g. FTSE100 dataset with $p = 20$ and the corresponding Table 3).

In the former cases, there is less chance of improving the performance of a portfolio provided by the CVaR-CC model; in the latter cases, the clusterization effect helps to achieve portfolios with a better out-of-sample performance. This observation is further illustrated in Figure 2, where we compare the weekly out-of-sample performance of two of our portfolios (SP500 with $p = 15$ Figure 2(a) and FTSE100 with $p = 20$ Figure 2(b), with a rebalancing allowed every 52 weeks) w.r.t. the ones provided by the CVaR-CC model.

5 Conclusion

In this paper we propose a novel framework for portfolio selection that combines the specific features of a clustering and a portfolio optimization techniques through the global solution of a hard Mixed-Integer Linear Programming problem. The idea of our approach is to overcome the classical two phases approach characterized by two distinct steps: cluster first and then selection. We show that our method is quite general since other risk measures as well as correlation measures can be adopted in our framework. The resulting MILP program is, in fact, *portable* in the sense that in model (4a)-(4g), only equations (4a)-(4d) depend on the specific risk measure adopted. Actually, constraints (4e)-(4g) are independent of such risk measure and they can be used in the formulation of MILP programs based on other measures (e.g., Mean Absolute Deviation, Gini's mean difference, minimax objective function etc...). Our model was tested on real financial datasets, compared to some benchmark models, and found to give good results in terms of realized profit. We also point out that the MILP program (4a)-(4g) can be efficiently solved at least for moderate sized problems.

To conclude, the results reported in this paper are encouraging and there is room for improving the optimization phase by providing an ad hoc heuristic approach for solving large size datasets. This is, in fact, one of our future lines of research.

Authors Contributions

Justo Puerto and Andrea Scozzari: Conceptualization, Methodology. Moisés Rodríguez-Madrena: Software Data curation. Andrea Scozzari: Writing- Original draft preparation. Justo Puerto: Supervision. Justo Puerto, Andrea Scozzari and Moisés Rodríguez-Madrena: Writing- Reviewing and Editing.

Acknowledgements

This research has been partially supported by Spanish Ministry of Economía and Competitividad/FEDER grants number MTM2016-74983-C02-01.

References

[1] Artzner, P., Delbaen, F., Eber, J. M., Heath, D.: Coherent Measures of Risk. Mathematical Finance 9, 203–228 (1999)

[2] Bellini, F., Di Bernardino, E.: Risk management with expectiles. The European Journal of Finance 23, 487–506 (2017)

[3] Benati, S.: Categorical data fuzzy clustering: an analysis of local search heuristics. Computers & Operations Research 35, 766–775 (2008)

[4] Benati, S., García, S.: A mixed integer linear model for clustering with variable selection. Computers & Operations Research 43, 280–285 (2014)

[5] Benati, S., García, S., Puerto, J.: Mixed integer linear programming and heuristic methods for feature selection in clustering. Journal of the Operational Research Society 69, 1379–1395 (2018)

[6] Benati, S., Rizzi, R.,: A mixed integer linear programming formulation of the optimal mean/Value-at-Risk portfolio problem. European Journal of Operational Research 176, 423–434 (2007)

[7] Beraldi, P., Bruni, M.E.: A clustering approach for scenario tree reduction: an application to a stochastic programming portfolio optimization problem. TOP 22, 934–949 (2014)

[8] Boginski, V., Butenko, S., Shirokikh, O., Trukhanov, S., Lafuente, J.G.: A network-based data mining approach to portfolio selection via weighted clique relaxations. Annals of Operations Research 216, 23–34 (2014)

[9] Bruni, R., Cesarone, F., Scozzari, A., Tardella, A.: On exact and approximate stochastic dominance strategies for portfolio selection. European Journal of Operational Research 259, 322–329 (2017)

[10] Bruni, R., Cesarone, F., Scozzari, A., Tardella, A.: Real-world datasets for portfolio selection and solutions of some stochastic dominance portfolio models. Data in Brief 8, 858–862 (2016)

[11] Cesarone, F., Scozzari, A., Tardella, A.: A new method for mean-variance portfolio optimization with cardinality constraints. Annals of Operations Research 205, 213–234 (2013)

[12] Cesarone, F., Tardella, A.: Equal Risk Bounding is better than Risk Parity for portfolio selection. Journal of Global Optimization 68, 439–461 (2017)

[13] Cesarone, F., Scozzari, A., Tardella, A.: Linear vs. quadratic portfolio selection models with hard real world constraints. Computational Management Science 12, 345–370 (2015)

[14] Chang, T.J., Meade, N., Beasley, J.E., Sharaiha, Y.M.: Heuristics for cardinality constrained portfolio optimisation. Computers & Operations Research 27, 1271–1302 (2000)

[15] Chopra, V. K., Ziemba, W. T.: The effect of errors in means, variances, and covariances on optimal portfolio choice. In Handbook of the Fundamentals of Financial Decision Making: Part I, pp. 365-373 (2013)

[16] Clemente, G. P., Grassi, R., Hitaj, A.: Asset allocation: new evidence through network approaches. Annals of Operations Research. https://doi.org/10.1007/s10479-019-03136-y (2019)

[17] Clemente, G. P., Grassi, R., Hitaj, A.: Smart network based portfolios. arXiv:1907.01274v1 [q-fin.PM] (2019)

[18] Corneil, D.G., Perl, Y.: Clustering and domination in perfect graphs. Discrete Applied Mathematics 9, 27–39 (1984)

[19] Corneil, D.G., Stewart, L.K.: Dominating sets in perfect graphs. Discrete Mathematics 86, 145–164 (1990)

[20] DeMiguel, V., Garlappi, L., Uppal, R.: Optimal versus naive diversification: How inefficient is the $1/n$ portfolio strategy? The Review of Financial studies 22, 1915–1953 (2007)

[21] Eftekhari, B., Pedersen, C.S., Satchell, S.E.: On the volatility of measures of financial risk: an investigation using returns from European markets. The European Journal of Finance 6, 18–38 (2000)

[22] Elton, E.J., Gruber, M.J., Spitzer, J.: Improved estimates of correlation coefficients and their impact on optimum portfolios. European Financial Management 12, 303–318 (2006)

[23] Ehrgott, M.: Multicriteria Optimization. Vol. 491. Springer Science & Business Media, Berlin Heidelberg (2005)

[24] Ehrgott, M.: A discussion of scalarization techniques for multiple objective integer programming. Annals of Operations Research 147, 343–360 (2006)

[25] García, S., Labbé, M., Marín, A.: Solving large p-median problems with a radius formulation. INFORMS Journal on Computing 23, 546–556 (2011)

[26] Garey, M.R., Johnson, D.S.: Computers and intractability: A guide to the theory of NP-Completeness. W.H. Freeman and Company, New York (1979)

[27] Hansen, P., Brimberg, J., Urosevic, D., Mladenovic, N.: Solving large p-median clustering problems by primal-dual variable neighborhood search. Data Mining and Knowledge Discovery 19, 351–375 (2009)

[28] Jegadeesh, N., Titman, S.: Profitability of momentum strategies: An evaluation of alternative explanations. The Journal of Finance 56, 699–20 (2001)

[29] Jobson, J. D., Korkie, B.: Estimation for Markowitz efficient portfolios. Journal of the American Statistical Association 75, 544-554 (1980)

[30] Kariv, O., Hakimi, L.: An Algorithmic Approach to Network Location Problems. II: The p-Medians. SIAM Journal on Applied Mathematics 37, 539–560 (1979)

[31] Köhn, H.F., Steinley, D., Brusco M.J.: The p-median model as a tool for clustering psychological data. Psycological Methods 15, 87–95 (2010)

[32] Konno, H., Yamazaki, H.: Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. Management science, 37, 519–531 (1991).

[33] Kullmann, L., Kertesz, J., Kaski, K.: Time-dependent cross-correlations between different stock returns: A directed network of influence. Physical Review E 66, 026125 (2002)

[34] Kullmann, L., Kertesz J., Mantegna, R.: Identification of clusters of companies in stock indices via Potts super-paramagnetic transitions. Physica A: Statistical Mechanics and its Applications 287, 412–419 (2000)

[35] Maillard, S., Roncalli, T., Telietche, J.: The properties of equally weighted risk contribution portfolios. The Journal of Portfolio Management 36, 60–70.

[36] Mansini, R., Ogryczak, W., Speranza, M. G.: Conditional value at risk and related linear programming models for portfolio optimization. Annals of operations research 152, 227–256 (2007)

[37] Mansini, R., Ogryczak, W., Speranza, M.G.: Linear and Mixed Integer Programming for Portfolio Optimization, Springer, Switzerland (2015)

[38] Mantegna, R.N.: Hierarchical structure in financial markets. The European Physical Journal B-Condensed Matter and Complex Systems 11, 193–197 (1999)

[39] Mantegna, R.N., Stanley, H.E.: An Introduction to Econophysics: Correlations and Complexity in Finance, Cambridge University Press, Cambridge (2000)

[40] Markowitz, H.: Portfolio selection. The Journal of Finance 7, 77–91 (1952)

[41] Merton, R.C.: On estimating the expected return on the market: An exploratory investigation. Journal of Financial Economics 8, 323–361 (1980)

[42] Mirchandani, P.B., Francis, R.L.: Discrete location theory, Wiley-Interscience Series in Discrte Mathematics and Optimization, New York (1990)

[43] Ng, R.T., Han, J.: Efficient and effective clustering methods for spatial data mining, in Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, San Francisco, CA, USA, 1994, Morgan Kaufmann Publishers Inc., pp. 144-155.

[44] J.-P. Onnela, J. P., Kaski, K., Kertesz J.: Clustering and information in correlation based financial networks. The European Physical Journal B 38, 353–362 (2004)

[45] Rockafellar, R. T., Uryasev, S.: Conditional value-at-risk for general loss distributions. Journal of banking & finance 26, 1443–1471 (2002)

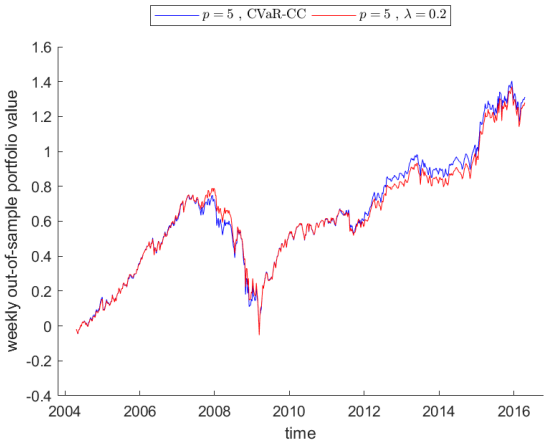
[46] Sharpe, W.F.: Mutual fund performance. Journal of Business 39, 119–138 (1966)

[47] Sharpe, W.F.: The sharpe ratio. The Journal of Portfolio Management 21, 49–58 (1996)

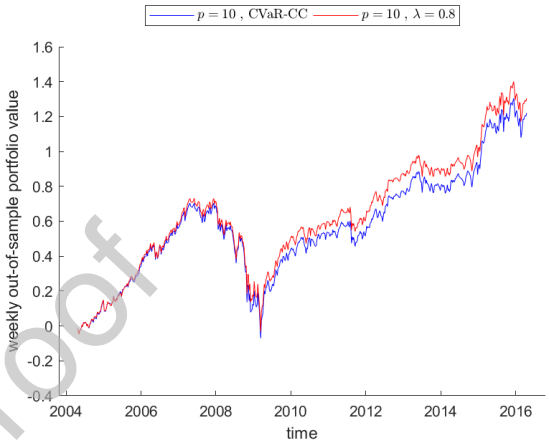
[48] Simaan, Y.: Estimation risk in portfolio selection: The Mean Variance Model and the Mean-Absolute Deviation model. Management Science 43, 1437–1446 (1997)

[49] Tola, V., Lillo, F., Gallegati, M., Mantegna, R.N.: Cluster analysis for portfolio optimization. Journal of Economic Dynamics and Control 32, 235–258 (2008)

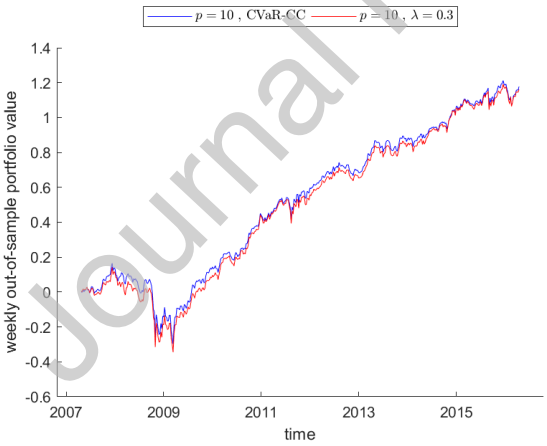
[50] Watts, D.J., Strogatz, S.H.: Collective dynamics of small-worldnetworks. Nature 393(6684), 440 (1998)



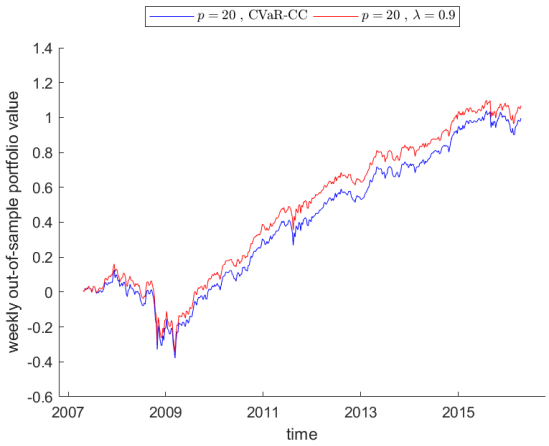
(a) Weekly out-of-sample values for EUROSTOXX50 $p = 5$ and $\lambda = 0.2$



(b) Weekly out-of-sample values for EUROSTOXX50 $p = 10$ and $\lambda = 0.8$

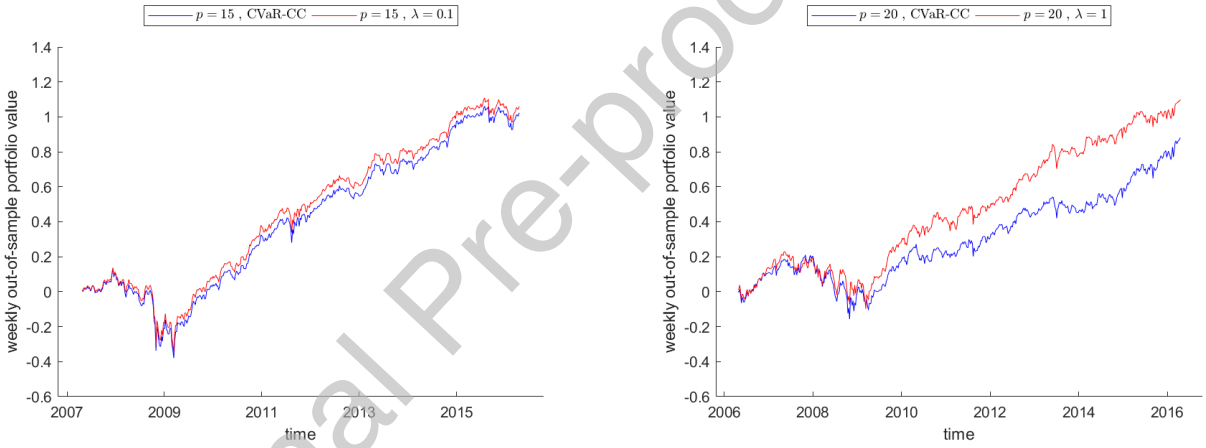


(c) Weekly out-of-sample values for SP500 $p = 10$ and $\lambda = 0.3$



(d) Weekly out-of-sample values for SP500 $p = 20$ and $\lambda = 0.9$

Figure 1: Weekly out-of-sample portfolio values for the EUROSTOXX50 and SP500 datasets for different values of p and λ.



(a) Weekly out-of-sample values for SP500 $p = 15$ and $\lambda = 0.1$

(b) Weekly out-of-sample values for FTSE100 $p = 20$ and $\lambda = 1$

Figure 2: Weekly out-of-sample portfolio values for the SP500 and FTSE100 datasets for different values of p and λ.