Enhanced indexation via chance constraints

Patrizia Beraldi¹ · Maria Elena Bruni¹

Received: 27 May 2019 / Revised: 10 June 2020 / Accepted: 1 August 2020 /
Published online: 1 September 2020
© The Author(s) 2020

Abstract

The enhanced index tracking (EIT) represents a popular investment strategy designed to create a portfolio of assets that outperforms a benchmark, while bearing a limited additional risk. This paper analyzes the EIT problem by the chance constraints (CC) paradigm and proposes a formulation where the return of the tracking portfolio is imposed to overcome the benchmark with a high probability value. Besides the CC-based formulation, where the eventual shortage is controlled in probabilistic terms, the paper introduces a model based on the Integrated version of the CC. Here the negative deviation of the portfolio performance from the benchmark is measured and the corresponding expected value is limited to be lower than a given threshold. Extensive computational experiments are carried out on different set of benchmark instances. Both the proposed formulations suggest investment strategies that track very closely the benchmark over the out-of-sample horizon and often achieve better performance. When compared with other existing strategies, the empirical analysis reveals that no optimization model clearly dominates the others, even though the formulation based on the traditional form of the CC seems to be very competitive.

Keywords Enhanced indexation · Chance constraints · Stochastic programming

1 Introduction

Over the last decades, index tracking (IT) has become an increasingly popular investment strategy all over the world. The aim is to create a portfolio of assets that replicates (tracks) the movements of a given financial index taken as benchmark. The difference between the performance of the replicating portfolio and the benchmark is generally referred to as tracking error and the problem entails the definition of a portfolio that minimizes this gap. On the basis of the specific

✉ Patrizia Beraldi
patrizia.beraldi@unical.it

¹ Department of Mechanical, Energy and Management Engineering, Università della Calabria,
Via P. Bucci, Cubo 41 C, Rende, Italy

tracking-error function used, different formulations have been proposed [see, for example, Mutunge and Haugland (2018) and the references therein].

The IT is traditionally referred to as a passive management strategy. In contrast, an active strategy aims at defining portfolios that achieve higher returns than the benchmark, generating an “excess of return”. Sometimes the term index-plus- α portfolio is used to indicate a portfolio that outperforms the benchmark by a given, typically small, value α . Since, in general, there is no guarantee of achieving a positive excess of return under every circumstance, the risk of underperforming the benchmark should be limited. Enhanced index tracking (EIT), or simply enhanced indexation, can be seen as an evolution of the IT relying on the idea of properly combining the strengths of both the passive and active management approaches (Canakgoz and Beasley 2009), by defining portfolios that outperform the benchmark while incurring a limited additional risk.

The definition of the IT strategies represents an active research area that has been receiving an increasing attention by both researchers and practitioners. Compared to the IT, the contributions dealing with the EIT problem are still limited and more recent. The first formalization of the problem, due to Beasley et al. (2003), dates back to 2003 and almost all the contributions appeared later than 2005.

Most of the proposed models rely on a backward perspective in that the tracking portfolios are built by considering as input data the historical observations in the hope that high accuracy in the past could guarantee the same result in the future. It is trivial to observe that a forward view would be advisable specially in periods of high financial turmoil as those experienced in recent years.

Among the few papers that analyze the problem under this latter perspective, we cite the contribution by Stoyan and Kwon (2010) who propose a two-stage stochastic programming formulation for the classical index tracking problem. In this paper, the alternative paradigm of Chance Constraints (CC, for short) introduced in Charnes and Cooper (1959) has been adopted with the aim of defining replicating portfolios whose performance can overcome the future benchmark return with a high probability level.

During the last decades, CC-based models have been applied to formulate several real-world problems (see, for example, Beraldi et al. 2012, 2015, 2017), where providing reliable solutions is considered a primary concern. Their use is actually not new in the field of portfolio optimization: the well-known Value at Risk measure, being a quantile, is modeled by a chance constraint and a large amount of papers on portfolio optimization can be “labeled” as CC-based models.

For the EIT problem the number of contributions is more limited and recent. A CC based formulation was proposed by Lejeune and Samatli-Pac (2013) where the variance of the benchmark return is forced to be below of a given threshold with a high probability. A deterministic equivalent reformulation is proposed under the assumption that the asset returns are represented by a stochastic factor model with the random factor following a normal distribution. More recently, by extending the reliable model proposed by Lejeune (2012), Xu et al. have presented (Xu et al. 2018) a sparse EIT model aimed at maximizing the excess return that can be attained with a high probability level. In the paper, the CC is approximated via the data driven

distributionally robust approach. Bruni et al. (2017) use the CC as relaxation of the Zero-order ϵ -stochastic dominance criteria.

In this paper, we present a stochastic formulation for the EIT problem where the CC is dealt under the assumption of discrete random variables. We point out that typically the CCs are formulated by considering continuous distributions, in particular the Gaussian one. For the considered application, this assumption might be inappropriate since most asset return distributions are leptokurtic or fat tailed. For the case of discrete random variables a deterministic equivalent reformulation of the stochastic problem can be obtained by the introduction of binary supporting variables, related to the realizations of the random variables, and a classical knapsack like constraint is used to limit the violation of the stochastic constraints. Depending on the total number of realizations used to model the uncertain parameters, the solution of the corresponding problem can be computationally demanding. In this paper, we empower the classical Branch and Bound method with a initial warm start solution determined by exploiting the specific problem structure.

Besides the CC model, we present another formulation based on the integrated version of the CC constraints (ICC). While the CC is used to control the probability of beating the benchmark and, as such, accounts for the “qualitative” side of the eventual shortage, the ICC provides a “quantitative” measure. In this case, the negative deviation of the tracking portfolio performance from the benchmark is quantified and the corresponding expected value is bounded to a given threshold. The ICC-based formulation presents some similarities with recent contributions as those considering the Conditional Value at Risk Measure [see, for example, Goel et al. (2018)] or that including second-order stochastic dominance (SSD) constraints [see, for example, Dentcheva and Ruszczyński (2003)] that, indeed, can be seen as a collection of the ICC.

The main contribution of this paper is to investigate the EIT problem by the CC paradigm assuming that the random variables follow a discrete distribution. The performance of the proposed models are also compared with other recent formulations on an out-of-sample basis.

The rest of the paper is organized as follows. Section 2 surveys the most recent literature on the EIT problem. Section 3 presents the proposed models, whereas Sect. 4 details the deterministic equivalent reformulations in the case of discrete random variables. Section 5 is devoted to the presentation of the computational experiments carried out to measure the performance of the proposed strategies also in comparison with other approaches proposed in the recent literature. Some conclusions and possible research directions are presented in Sect. 6.

2 Related literature

The growing popularity of the enhanced index funds, experienced both in mature and emerging markets (Parvez and Sudhir 2005; Weng and Wang 2017), has pushed the academic community towards the design of quantitative tools to support investors. While the scientific literature on the IT is rather extensive [see, for example, Sant’Anna et al. (2017) for a recent overview], the number of contributions on the

EIT problem is lower but steadily increasing. The EIT problem was originally introduced by Beasley et al. (2003) and an overview of the early literature can be found in Canakgoz and Beasley (2009). In what follows, we review the most recent contributions focusing on the main measures used to control the risk of underperforming the benchmark.

A first group of contributions adopt the variance (or the standard deviation) as risk measure. For example, in (de Paulo et al. 2016) De Paulo et al. consider the variance computed on the difference between the replicating portfolio return and the benchmark and propose a mean-risk approach. Standard deviation has been used by Wu et al. (2007), where the authors consider a bi-objective function, dealt by a goal programming technique, with the term to be maximized set equal to the excess portfolio return. More recently, Cesarone et al. have proposed (Cesarone et al. 2019) a novel approach aimed at finding a Pareto optimal portfolio that maximizes the weighted geometric mean of the difference between its risk and gain and those of a suitable benchmark index. In the experiments, the authors use as risk measure the standard deviation of the replicating portfolio. A mean-risk structure has been also adopted in Li et al. (2011) where the tracking error is expressed as downside standard deviation. The tracking error variance has been used in the very recent contribution (Gnägi and Strub 2020) where the authors propose a model aimed at minimizing a quadratic function of the covariances between the returns of the stocks, the weights of the stocks in the composing portfolio and the weights of the stocks in the benchmark. The authors showed that minimizing this risk measure may provide better performance in terms of the out-of-sample tracking error when compared with other strategies. The main limitation of this approach is related to necessity of knowing the actual composition of the index that is often not available for the existing test instances.

Absolute deviation related measures have been used in several other contributions. Koshizuka et al. proposed (Koshizuka et al. 2009) a model aimed at minimizing the tracking error from an index-plus- α portfolio, choosing among the portfolios with a composition highly correlated with the benchmark. Two alternative measures of the tracking error have been considered: one based on the absolute deviation and the other on the downside absolute deviation. In Valle et al. (2014) the authors studied an absolute return portfolio problem and propose a three-stage solution approach. The maximum downside deviation of the replicating portfolio compared to the benchmark was used by Bruni et al. (2015) in a bi-objective approach.

More recently, the Conditional Value at Risk (CVaR) measure has been used to model risk in the EIT problem. For example, in Goel et al. (2018) the authors consider two variants of the CVaR: the two tails CVaR and the mixed one. While the former measure represents the weighted sum of the left tails (worst scenarios) and right tails (best scenarios), the latter is the weighted sum of the CVaR calculated for different confidence levels. The mixed CVaR has been also used by Guastaroba et al. (2017) in the framework of risk-reward ratio. The contribution is similar to the model proposed by the same authors in Guastaroba et al. (2016) where they consider the maximization of the Omega Ratio in the standard and the extended forms.

Stochastic dominance (SD) criteria, in different forms, have been also applied to control the risk of underperforming the benchmark. The basic aim is to create

replicating portfolios whose random return stochastically dominates the return of the benchmark. One of the first EIT model based on the SD has been proposed by Kuosmanen (2004), who applied the First-order Stochastic Dominance (FSD) and Second Order SD rules. Later on, in Luedtke (2008) presented compact linear programming formulations where the objective is to maximize the portfolio expected return with SSD constraints over the benchmark. More recently, Roman et al. (2013) applied a SSD strategy to construct a portfolio whose return distribution dominates the benchmark ones. We note that typically the solution of a problem including SSD constraints is computationally demanding since the number of constraints to include is of the order of the number of scenarios squared. Moreover, the SSD conditions are judged by many agents as excessively demanding because the most extreme risk averse preferences are taken into account as well. For this reason, some forms of relaxation of the SD have been proposed. Sharma et al. (2017) used underachievement (surplus) and overachievement (slack) variables and control the relaxed SSD condition by imposing bounds on the ratio of the total underachievement to the sum of total underachievement and overachievement variables. More recently, Bruni et al. (2017) applied a Zero-order ϵ -SD rule and its cumulative version. Higher order SD rules, that can be viewed as relaxation of SSD as well, have been applied by Post and Kopa (2017).

As final contributions, we mention the papers, more related to our work, that use the chance constraints to control the risk. In Lejeune (2012) propose a stochastic model aimed at maximizing the probability to obtain excess return of the invested portfolio with respect to the benchmark. Recently, an extension of the previous mentioned paper has been proposed by Xu et al. (2018), where the authors present a sparse enhanced indexation model aimed at maximizing the excess of return that can be achieved with a high probability. The proposed problem is dealt by a distributionally robust approach. In the same stochastic stream, we cite Lejeune and Samatli-Pac (2013) where the CC is used to impose the variance of the index fund return to be below a threshold with a large probability. A deterministic equivalent formulation, exploiting the classical standardization technique, is provided under the assumption that the asset returns are represented by a stochastic factor model, with the random factor following a Normal distribution.

Our CC formulation share with these last mentioned papers the choice of the modeling paradigm: the CC is used to guarantee that the performance of the replicating portfolio can beat the benchmark with a given probability level. Differently, the CC is dealt under the assumption of random variables with discrete distribution represented by a set of scenarios each occurring with a given probability level. The formulation based on the integrated form of the CC controls the expected shortage and, as such, shares some similarities with CVaR based models and with the SSD ones, since SSD can be seen as a family of ICC. In the paper, the CC based formulations are compared each other and with other recent approaches proposed for the EIT problem. The extensive computational phase can be seen as another contribution of the paper.

3 The EIT problem under chance constraints

We consider the problem of a portfolio manager who wants to determine a portfolio that outperforms a benchmark, represented by the rate of return of a market index eventually increased by a given value. A buy and hold strategy is supposed to be adopted in that the defined portfolio is kept until the end of a given investment horizon. Let $J = \{1, 2, \dots, N\}$ denote the set of the index constituents. A portfolio is identified by specifying the fraction x_j of capital invested in every asset $j \in J$. No short sales are allowed (x_j are non negative decision variables) and the consistency constraint $\sum_{j \in J} x_j = 1$ is imposed. We denote by $\mathcal{X}$ the feasible set determined by these basic constraints. Each asset j , as well as the benchmark, generate an uncertain return modeled as random variables, $\tilde{r}_j$ and $\tilde{\beta}$, defined on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Under this assumption, the EIT problem becomes a stochastic optimization problem:

$$\begin{aligned} & \max f(\widetilde{R}_p, \tilde{\beta}) \\ & \widetilde{R}_p \geq \tilde{\beta} \\ & x \in X \end{aligned} \quad (1)$$

where $\widetilde{R}_p = \sum_{j \in J} \tilde{r}_j x_j$ denotes the random portfolio return.

The way in which the feasibility and optimality conditions of problem (1) are redefined depends on the approaches used to deal with the randomness. In the proposed formulation, the objective function is dealt by considering the expected value operator $\mathbb{E}[\cdot]$, whereas the stochastic constraint is treated by adopting the CC paradigm:

$$\mathbb{P}(\widetilde{R}_p \geq \tilde{\beta}) \geq (1 - \gamma) \quad (2)$$

Roughly speaking, the violation of the stochastic constraint is allowed provided that it occurs with a low probability value denoted by the parameter γ in (2). This value represents the allowed tolerance level and is chosen by the decision maker on the basis of his risk perception. Low values of γ (eventually 0) are used to model very risk averse positions: the replicating portfolios should beat the benchmark under every possible circumstances that might occur in the future. A part from the feasibility issue, a natural consequence of this pessimistic choice is the over conservativeness of the suggested investment policies. Higher values of γ , on the contrary, may provide portfolios with higher potential returns, as long as the investor is willing to accept the risk of underperforming the benchmark.

Dating back to the 60th (Charnes and Cooper 1959), the CC still represents a challenging topic in optimization. All the existing approaches for solving the CC-based problems rely on the derivation of deterministic equivalent reformulations that, in turn, depend on the nature of the random variables. In the case of Normal random variables, a deterministic equivalent reformulation can be easily derived by applying standard normalization techniques. In particular, it is easy to show that (2) can be rewritten as

$$\sum_{j \in J} \bar{r}_j x_j \geq \bar{\beta} - \phi^{-1}(\gamma) \sqrt{\sum_{j \in J} \sigma_j^2 x_j^2 + \sum_{j \in J} \sum_{i \in J, i \neq j} \sigma_{ji} x_j x_i - 2 \sum_{j \in J} \sigma_{j\beta} x_j + \sigma_\beta^2} \quad (3)$$

Here $\bar{r}_j$ and $\bar{\beta}$ represent the expected rate of return of the assets and the benchmark, respectively, whereas σ_{ji} , $\sigma_{j\beta}$, σ_β^2 denote the covariance between the return of asset j and i , between the asset j and the benchmark return, and the variance of the benchmark return, respectively. Finally, $\phi^{-1}(\gamma)$ is the γ -quantile of the normal standard distribution. The deterministic reformulation highlights that the expected return of the replicating portfolio is required to exceed the expected benchmark return augmented by a penalty term (i.e. $\phi^{-1}(\gamma)$ is a negative term) that increases as γ decreases. We note that a reformulation similar to (3) can be derived when the distribution probability $\mathbb{P}$ is unknown, provided that the first two moments of the distribution are known. In this case the term multiplying the standard deviation is $\sqrt{\frac{1-\gamma}{\gamma}}$.

In this paper, we deal with the CC assuming that the random variables follow a discrete distribution. This hypothesis allows to overcome the shortcomings related to the choice of the Normal distribution that is inappropriate in the case of distributions exhibiting fat tails. Discrete realizations can be derived by using specific scenario generation techniques [see, for example, Beraldi et al. (2010) and the references therein] or available historical observations. The following section presents the deterministic equivalent reformulation under this assumption.

It is worthwhile noting that the CC (2) allows to control the probability of beating the benchmark, thus accounting for the “qualitative” side of the eventual shortage. Even though this probability is limited, the shortage, if experienced, could be high. Under this respect, the integrated version of the CC provides a way to keep the expected shortage under control. According to this paradigm, the stochastic constraint (2) is written as

$$\mathbb{E}[(\tilde{\beta} - \tilde{R}_p)_+] \leq \epsilon \quad (4)$$

where $(\tilde{\beta} - \tilde{R}_p)_+$ denotes the random shortage of the replicating portfolio return with the respect to the benchmark. In this case, the probability distribution is used to measure the expected magnitude of the shortage and both quantitative and qualitative aspects are jointly accounted for. The parameter ϵ is defined by the end-user on the basis of the expected shortage that he is willing to accept.

4 The proposed formulations

This section presents the CC-based reformulations under the assumption of discrete random variables. We denote by $\mathcal{S} = \{1, \dots, S\}$ the set of scenarios, i.e. realizations that the random parameters can take, and by π_s the probability of occurrence.

4.1 The CC-based model

Under this assumption, constraint (2) can be rewritten in the following disjunctive form:

$$\Gamma(1 - \gamma) = \bigcup_{K \in \mathcal{K}} \bigcap_{s \in K} \{R_p^s \geq \beta^s\}. \quad (5)$$

Here the set $\mathcal{K}$ is defined as

$$\mathcal{K} = \{K | K \subseteq \mathcal{S}, \sum_{s \in K} \pi^s \geq 1 - \gamma\} \quad (6)$$

whereas R_p^s and β^s denote the s -realization of the portfolio return and the benchmark, respectively. A natural approach to rewrite the constraint (5) is by “big-M” reformulation:

$$\sum_{j \in J} r_{js} x_j + M z_s \geq \beta^s \quad \forall s \quad (7)$$

$$\sum_{s \in \mathcal{S}} \pi_s z_s \leq \gamma \quad (8)$$

$$z_s \in \{0, 1\} \quad (9)$$

where M is a real number large enough to ensure that if the binary variable z_s takes value 1, then (7) is not active, the opposite when z_s is 0. Constraint (8) is a knapsack restriction that limits to γ the violation of the scenario dependent constraints.

The complete CC-based reformulation of the EIT problem is a mixed integer problem with the following structure:

$$\max \sum_{j \in J} \bar{r}_j x_j \quad (10)$$

$$(7) - (9) \quad (11)$$

$$x \in \mathcal{X} \quad (12)$$

Depending on the cardinality of the scenario set and by the imposed risk level γ , the solution of the reformulated problem can be computational demanding. Let us assume, for example, that each scenario s has the same probability of occurrence, i.e. $\pi_s = 1/S$. In this case, constraint (8) can be rewritten as:

$$\sum_{s \in \mathcal{S}} z_s \leq q \quad (13)$$

with q equal to $\lfloor \gamma S \rfloor$. Let us denote by $\mathcal{S}_q$ the whole set of feasible solutions of constraint (13). Each solution is associated with a subset $\mathcal{I}_q$ of q scenarios. A trivial

solution approach to solve the CC-based problem would rely on an enumeration scheme: for each element of $\mathcal{S}_q$, solve the following linear programming problem:

$$\max \sum_{j \in J} \bar{r}_j x_j \quad (14)$$

$$R_p^s \geq \beta^s \quad \forall s \in \mathcal{S} - \mathcal{I}_q \quad (15)$$

$$x \in \mathcal{X} \quad (16)$$

The optimal solution of the original problem will be the best among the determined ones. It is evident that such an approach is prohibitive even for moderate size of the scenario set, since the number of sets $\mathcal{S}_q$ is equal to $\binom{S}{q}$.

Solution approaches exploiting the specific problem structure have been proposed in the last decades (e.g. Beraldi and Bruni 2010, Bruni et al. 2013, Beraldi and Ruszczyński 2005). In this paper, we use the classical Branch and Bound approach implemented in commercial solvers, empowered with a procedure for determining an initial incumbent solution. The basic idea, reported in the following scheme, relies on the use of the Lagrangian multipliers λ_s associated with the scenario based constraints. Since a positive value of the multiplier represents the sensitivity of the optimal objective function value to variations in the s -th constraint, we adopt a locally-best choice by removing the q constraints associated with the largest values. Finally, the reduced problem containing the promising scenarios is solved as integer and the corresponding objective function value is imposed as initial incumbent value. The basic scheme is reported in the following.

- **Initialization.** Create the supporting set $\mathcal{R}_q$ initially containing all the scenarios.
- **Step 1.** Solve model (14)–(16) by imposing the scenario constraints for all $s \in \mathcal{R}_q$.
- **Step 2.** If the problem is feasible go to Step 3 otherwise go to Step 4.
- **Step 3.** Let λ_s be the Lagrangian multipliers associated with the scenario constraints. Sort these values in non increasing order. Compose the set $\mathcal{I}_q$ in (15) by using the corresponding first q scenarios. Solve model (14)–(16) and set the initial incumbent equal to the corresponding objective function value. Go to Step 5.
- **Step 4.** Randomly select a scenario l and update the set $\mathcal{R}_q = \mathcal{R}_q - l$. Go to Step 1.
- **Step 5.** Invoke the classical Branch & Bound algorithm.

We note that when the model solved at Step 1 is infeasible, an iterative procedure is applied consisting in removing a scenario constraint at time. Other rules for selecting a given number of q scenarios out of the S can be applied (as for example, a simple, random choice). However, we have found that simple initialization allows a nice reduction of the computational effort, when used as mip starter in a classical Branch & Bound algorithm.

4.2 The ICC-based model

In the case of the ICC the reformulation is less involved and relies on the introduction of continuous scenario dependent variables y_s defined as $\max(0, \beta^s - R_p^s)$. More specifically, constraint (4) can be written as:

$$y^s \geq \beta^s - \sum_{j \in J} r_{js} x_j \quad \forall s \quad (17)$$

$$\sum_{s \in \mathcal{S}} \pi_s y_s \leq \epsilon \quad (18)$$

$$y_s \geq 0 \quad \forall s \quad (19)$$

It is worthwhile observing that the ICC paradigm is related to other approaches recently proposed for the EIT problem. An evident connection is with the cumulative ϵ -ZSD introduced in Bruni et al. (2017). While the ϵ -ZSD imposes that under every scenario s the underperformance with respect to the benchmark β^s is limited by ϵ , by the cumulative condition, the aggregated loss is taken into account. Instead of considering the bound ϵ as a parameter, the authors consider it as decision variable to be minimized. In our formulation, we still consider an aggregated loss, but we also take the scenario probability in the expected value.

Finally, we note the connection between the ICC and the SSD constraints. We recall that a random variable $\tilde{R}_p$ dominates in the second order a random variable $\tilde{\beta}$ if

$$F_{\tilde{R}_p}^2(\eta) \leq F_{\tilde{\beta}}^2(\eta) \quad \forall \eta \in R \quad (20)$$

where $F^2(\cdot)$ denotes the second performance function, that can be also expressed as

$$F_V^2 = \mathbb{E}[(\eta - V)_+]$$

with V representing a generic random variable.

Under the assumption of discrete distributions, constraint (20) can be written as:

$$\mathbb{E}[(\beta_t - \tilde{R}_p)_+] \leq \mathbb{E}[(\beta_t - \tilde{\beta})_+] \quad \forall t \quad (21)$$

Thus, the SSD constraints can be written as a collection of ICC:

$$\sum_{j \in J} r_{js} x_j + l_{ts} \geq \beta_t \quad \forall t, s \quad (22)$$

$$\sum_{s \in \mathcal{S}} \pi_s l_{ts} \leq v_t \quad \forall t \quad (23)$$

$$l_{ts} \geq 0 \quad \forall t, s \quad (24)$$

where $v_t = \mathbb{E}[(\beta_t - \tilde{\beta})_+]$.

5 Computational experiments

This section is devoted to the presentation and discussion of the computational experiments carried out with the aim of empirically assessing the effectiveness of the proposed models. We have considered the data sets used in Guastaroba et al. (2016), consisting of two groups of instances, namely the *GMS* and the *ORL*. For all the instances, that are available at the website (<http://or-brescia.unibs.it>), the number of scenarios corresponds to two-years weekly observations and is equal to 104. The performance of the generated portfolios are evaluated on an out-of-sample basis, by measuring their return in the 52 weeks following the date of portfolio selection. Table 1 reports the main characteristics of the tested instances: the name of the instance, the benchmark, the number of available assets and of scenarios. The *GMS* set includes four instances created to span four different market trends. For example, the instance referred to as *GMS-UU* is characterized by an increasing trend of the market (i.e. the market is moving up) in the in-sample periods as well as in the out-of sample periods. Similar considerations are valid for the other *GMS* instances.

The *ORL* set is generated from 8 benchmark instances for the index tracking problem that consider different stock market indices.

Besides using historical data as realizations of the random asset returns, additional tests have been carried out by adopting a MonteCarlo scenario generation technique.

In particular, we have assumed that random asset prices are modeled by a correlated Brownian Motion (the market index is dealt as an additional asset in our generation). Starting from the historical observations, we have determined for each asset i the expected returns μ_i , and the Variance Covariance matrix with elements σ_i^2 denoting the variance of the return of asset i and σ_{ij} representing the covariance between the returns of asset i and j . For every asset i the following formula has been applied:

Table 1 Data set

Data set	Instance	Benchmark	N	S
GMS	GMS-UU	FTSE100	100	104
	GMS-UD	FTSE100	100	104
	GMS-DU	FTSE100	100	104
	GMS-DD	FTSE100	100	104
ORL	ORL-IT1	Hang Seng	31	104
	ORL-IT2	Dax100	85	104
	ORL-IT3	FTSE100	89	104
	ORL-IT4	S&P100	89	104
	ORL-IT5	Nikkei225	225	104
	ORL-IT6	S&P500	457	104
	ORL-IT7	Russel2000	1318	104
	ORL-IT8	Russel3000	2151	104

$$r_i = (\mu_i - (\sigma_i^2)/2) + \sigma_i \sum_{k=1}^i C_{ik} \epsilon_k$$

where C_{ik} are the coefficients of the Cholesky's matrix of the variance covariance matrix and $\epsilon_k \sim N(0,1)$. By applying the MonteCarlo simulation different scenario sets of increasing cardinality have been determined. In particular, in our tests we have considered instances with 150, 250 and 450 scenarios. These new instances are referred by adding to the original name the scenario number. Thus, for example ORL-IT8-250 refers to the version of the original instance where 250 scenarios are generated by the MonteCarlo Simulation technique. For all the tested instances, the same probability of occurrence has been associated with the scenarios.

The results presented here after mainly refer to the set of test instances available in literature, i.e. scenarios are determined by considering historical observations. The motivation of this choice is related to the possibility of making the results easily reproducible. The last subsection is devoted to the presentation and discussion of the the results obtained by considering the MonteCarlo simulation technique.

All the models have been implemented in GAMS 24.7.4 and solved by CPLEX 12.6.1 on a C Intel Core I7 (2.5 GHz) with 8 GB of RAM DDR3. The solution of the CC formulations has been carried out by applying the specialized Branch & Bound approach described in Sect. 4.1. We note that the solution time increases as higher γ values are considered. For those instances, the application of the specialized approach allows to achieve a reduction of the solution time around 20%. Anyhow, even for the larger instances the solution times are still limited. As for the ICC formulation, that entails the solution of a LP problem, the required computational times are much lower (around few seconds).

5.1 Numerical results

Different computational experiments have been carried out to evaluate the performance of the proposed CC models.

Figure 1 reports the expected portfolio return as function of the risk aversion level γ . The numbers shown next to the graph denote the standard tracking error (computed by considering the deviation below the benchmark).

The results refer to the test case GMS-DD, but a similar behavior in terms of risk-return trade-off has been observed for all the other tested instances.

As evident, when lower values of γ are considered, the replicating portfolios show lower expected returns and, as expected, lower tracking errors. As the γ value is increased, the performance improves (the problem is less constrained), but the tracking error slightly worsens. By varying the γ values, portfolios with different performance can be obtained, thus providing the decision maker with a wide range of solutions to choose from according to the investor's risk attitude. Very risk adverse investor may be willing to sacrifice some return as long as "safer" solutions are guaranteed.

The results on the portfolio composition confirm that the lower the γ value the larger the number of selected assets. For example, for the test case introduced above,

Fig. 1 Efficient frontier for the GMS-DD instance

the number of assets ranges between 38 and 20 when passing from 0.01 to 0.10. The same trend has been observed for all the tested instances that present a limited number of assets included in the replicating portfolios. Similar results have been obtained for the ICC model, that on the whole shows slighter worse performance.

Additional experiments have been carried out to evaluate the realized performance of the CC formulations on an out-of-sample basis. The following figures report the cumulative return of the benchmark and the replicating portfolios as function of γ for the GMS instances.

The results clearly show that the CC portfolios mimic closely the behavior of the market over the entire out-of-sample horizon: cumulative returns of the benchmark and the replicating portfolios jointly increase and/or decrease. Moreover, for some periods the replicating portfolios outperform the benchmark. Analyzing the results in more details, we may observe that when the market is down both in the in-sample and out-of-sample periods (GMS -DD instance) the best results have been obtained when considering higher values of γ , as shown in Fig. 2.

This behavior can be explained observing that when the market experiences a negative trend, tracking the benchmark as close as possible is not a winning strategy in the attempt of gaining more. On the contrary, when the market is up (see Fig. 3), the best strategy seems to be the faithful replication of the benchmark obtained by choosing lower values of γ .

When the market has a mixed trend, lower values of γ are preferable as shown in Figs. 4 and 5 that refer to the GMS-UD and GMS-DU instances.

For the ORL data set, the results (not reported here for the sake of brevity) are similar and the best performance are obtained for small values of γ .

The results reported hereafter refer to the out-of-sample analysis for the ICC formulation. The following Figs. 6 and 7 show the cumulative returns for the GMS - DD and GMS - UD instances, respectively. Similar behavior has been observed for the other tested instances. As evident, the replicating portfolios track closely the ex-post behavior of the benchmark. In some cases (see Fig. 6) the return patterns seem to overlap, in some others the tracking portfolio provides better

Fig. 2 CC formulation: cumulative realized returns as function of γ

Fig. 3 CC formulation: cumulative realized returns as function of γ

performance when compared to the market index (see Fig. 7). In all the cases, the results seem to be quite satisfactory.

5.2 Comparison with other strategies

Additional experiments have been carried out to provide a comparison between the CC based formulations and other strategies recently proposed for the EIT problem. For the CC model, since difference performance have been obtained for different values of γ , we report the average results. As basis of comparison we have considered the model presented in Guastaroba et al. (2016), referred as omega, where the

Fig. 4 CC formulation: cumulative realized returns as function of γ

Fig. 5 CC formulation: cumulative realized returns as function of γ

objective function is the maximization of the extended omega ratio (the value of α has been set to 0) and the SSD formulation (22)–(24). The analysis has been carried out by computing some performance measures typically used in the portfolio optimization literature, namely the average return, the Sortino Ratio and the number of weeks, divided by 52 and in percentage, that the replicating portfolio beats the benchmark. All the values have been computed on the out-of-sample time horizon and for all the measures, larger values are preferred.

The following Tables report the values of the different measures for the tested models (Tables 2, 3). The best values recorded for all the measures are reported in bold.

Fig. 6 ICC formulation: cumulative realized returns

Fig. 7 ICC formulation: cumulative realized returns

Looking at the overall results, we may observe that there is no formulation that dominates the others according to all the computed measures and for all the tested instances. Table 4 reports the number of times that a given approach provides the best result with the respect to a given measure.

The results show that the CC formulation provides better results when compared with the ICC one. The SSD formulation presents satisfactory results that are better than those provided by the ICC formulation. Anyhow, it is no possible to establish the winning or losing model. More interestingly, it is possible to confirm the efficacy of the proposed models in supporting investment decisions suggesting strategies able to outperforms the benchmark ex-post.

Table 2 Out-of-sample results for the GMS instances

Instance	Approach	Aver. return	Sortino	Ntimes
GMS-DD	CC	−0.0026	0.1807	59.61
	ICC	−0.0055	−0.1136	44.23
	Omega	−0.0049	0.0216	50
	SSD	−0.0034	0.0664	50
GMS-DU	CC	0.0066	0.2225	61.53
	ICC	−0.0058	−0.0796	50
	Omega	0.0055	−0.0216	51.92
	SSD	0.0127	0.2286	67.30
GMS-UD	CC	0.0001	0.3807	57.69
	ICC	−0.0185	−0.9648	50
	Omega	0.0005	0.5091	67.30
	SSD	−0.0087	−0.1994	42.30
GMS-UU	CC	0.0073	0.0756	58
	ICC	0.0307	0.0315	53.8
	Omega	0.0071	−0.0252	59.61
	SSD	−0.0022	−0.3024	42.30

5.3 Solution of larger instances

This section is devoted to the presentation and discussion of the results obtained when considering a larger number of scenarios generated by a MonteCarlo simulation technique. As expected, the quality of the scenarios affects the realized performance. The following Table reports the results obtained for the test instance ORL-IT8 as function of the scenario number (Table 5). Looking at the results, we may notice that better performance can be achieved when the Montecarlo simulation is adopted. For example, considering a high level of risk aversion ($\gamma = 0.99$) the realized average return passes from 0.0072 when using historical observation (104 scenarios) to 0.050 when the scenarios (150) are generated by to the MonteCarlo simulation. The improvement in terms of Sortino ratio and number of times the portfolio outperforms the benchmark is also impressive, especially when considering a high level of risk aversion. We may notice that the performance seems to deteriorate when passing from 250 to 450 scenarios. Anyhow, even for in this case the results are much better than those achieved when historical data are used. If we look at the values averaged considering the different sizes of the scenario set, it appears evident that the best results are obtained considering a γ value equal to 0.01, confirming that keeping a prudent attitude pays on the long run. We finally notice that the same considerations drawn for the ORL-IT8 test can be drawn for all the other instances. The whole set of results is not reported here for the sake of brevity.

Table 3 Out-of-sample results for the ORL instances

Instance	Approach	Aver. Return	Sortino	Ntimes
ORL-IT1	CC	−0.0030	0.0320	51.92
	ICC	−0.9240	−0.0784	48.07
	Omega	− 0.0027	0.2389	59.61
	SSD	0.0043	0.4506	48.0769
ORL-IT2	CC	0.0002	0.0737	51.92
	ICC	0.0003	0.1104	59.61
	Omega	0.0002	0.0782	61.52
	SSD	0.0004	0.0219	50
ORL-IT3	CC	− 0.0011	0.0255	57.60
	ICC	−0.0014	−0.0316	50
	Omega	−0.0015	−0.0360	50
	SSD	− 0.0006	0.0683	53.84
ORL-IT4	CC	0.0011	0.0339	57.69
	ICC	0.0008	−0.0683	44.23
	Omega	0.0008	−0.0337	44.22
	SSD	−0.0006	−0.1312	41.23
ORL-IT5	CC	−0.0037	−0.2930	48.07
	ICC	−0.0036	−0.3056	44.23
	Omega	−0.0035	−0.2842	44.22
	SSD	−0.0067	−0.1951	41.23
ORL-IT6	CC	0.0107	0.4152	67.94
	ICC	0.0045	0.0968	44.23
	Omega	0.0047	0.1408	44.22
	SSD	0.0004	−0.0470	41.23
ORL-IT7	CC	0.0061	0.1818	67.30
	ICC	0.0045	0.0968	44.23
	Omega	0.0047	0.1408	44.22
	SSD	0.0004	−0.0470	41.23
ORL-IT 8	CC	0.0064	0.14088	50
	ICC	0.0071	0.1623	48.07
	Omega	0.0077	0.2360	51.91
	SSD	0.0011	−0.0541	50.51

Table 4 Comparison of the different strategies

Approach	Aver. Return	Sortino	Ntimes
CC	5	4	6
ICC	1	2	1
Omega	2	2	4
SSD	3	4	1

Table 5 Out-of-Sample results for the ORL-IT8 instance as function of the scenario number

γ	Scenarios	Aver. Return	Sortino	Ntimes
0.01	104	0.0072	0.2214	48.07
	150	0.0501	3.4067	80.76
	250	0.0510	3.382	84.61
	450	0.0378	2.0176	86.53
	Average	<i>0.0460</i>	<i>2.9375</i>	<i>83.97</i>
0.05	104	0.0070	0.1561	51.92
	150	0.0501	3.3880	84.61
	250	0.0507	1.7466	75
	450	0.0220	0.3147	57.69
	Average	<i>0.0409</i>	<i>1.8164</i>	<i>72.43</i>
0.10	104	0.0049	0.0450	50
	150	0.0378	2.0169	86.53
	250	0.0220	0.3147	57.69
	450	0.0120	0.1124	48.07
	Average	<i>0.0239</i>	<i>0.8147</i>	<i>64.10</i>

γ denotes the risk a version

6 Conclusions

The enhanced index tracking problem represents a challenging problem that has been receiving an increasing attention by the scientific community in the last decades. In this paper, the problem is addressed by applying the machinery of chance constraints and two formulations, based on the basic and the integrated form of the stochastic constraint are proposed. Extensive computational experiments are carried out on different benchmark instances. Both the proposed formulations suggest investment strategies that track very closely the benchmark over the out-of-sample horizon and often achieve better performances. When compared with other existing strategies, the empirical analysis reveals that no optimization model clearly dominates the others in the out-of-sample analysis, even though the chance constrained formulation seems to be very competitive. Additional tests have been carried out to evaluate the impact of the quality of the scenarios used as input data in the optimization models. We have empirically found that when scenarios are generated by applying a MonteCarlo simulation technique the realized returns are superior than those achieved when considering historical observations. The results confirm that the adoption of a forward perspective is preferable since high accuracy in the past seldom guarantees the same result in the future.

Funding Open access funding provided by Università della Calabria within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative

Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Beasley J, Meade N, Chang T-J (2003) An evolutionary heuristic for the index tracking problem. *Eur J Oper Res* 148(3):621–643
- Beraldi P, Bruni ME (2010) An exact approach for solving integer problems under probabilistic constraints with random technology matrix. *Ann Oper Res* 177(1):127–137
- Beraldi P, Ruszczyński A (2005) Beam search heuristic to solve stochastic integer problems under probabilistic constraints. *Eur J Oper Res* 167(1):35–47
- Beraldi P, Simone FD, Violi A (2010) Generating scenario trees: a parallel integrated simulation optimization approach. *J Comput Appl Math* 233(9):2322–2331
- Beraldi P, Bruni M, Violi A (2012) Capital rationing problems under uncertainty and risk. *Comput Optim Appl* 51(3):1375–1396
- Beraldi P, Bruni M, Laganá D, Musmanno R (2015) The mixed capacitated general routing problem under uncertainty. *Eur J Oper Res* 240(2):382–392
- Beraldi P, Violi A, Bruni M, Carrozzino G (2017) A probabilistically constrained approach for the energy procurement problem. *Energies* 10(12)
- Bruni ME, Beraldi P, Laganá D (2013) The express heuristic for probabilistically constrained integer problems. *J Heuristics* 19(3):423–441
- Bruni R, Cesarone F, Scozzari A, Tardella F (2015) A linear risk-return model for enhanced indexation in portfolio optimization. *OR Spectrum* 37(3):735–759
- Bruni R, Cesarone F, Scozzari A, Tardella F (2017) On exact and approximate stochastic dominance strategies for portfolio selection. *Eur J Oper Res* 259(1):322–329
- Canakgoz N, Beasley J (2009) Mixed-integer programming approaches for index tracking and enhanced indexation. *Eur J Oper Res* 196(1):384–399
- Cesarone F, Lampariello L, Sagratella S (2019) A risk-gain dominance maximization approach to enhanced index tracking. *Finance Res Lett* 29:231–238
- Charnes A, Cooper WW (1959) Chance-constrained programming. *Manag Sci* 6(1):73–79
- de Paulo WL, de Oliveira EM, do Valle Costa OL (2016) Enhanced index tracking optimal portfolio selection. *Finance Res Lett* 16:93–102
- Dentcheva D, Ruszczyński A (2003) Optimization with stochastic dominance constraints. *SIAM J Optim* 14(2):548–566
- Gnägi M, Strub O (2020) Tracking and outperforming large stock-market indices. *Omega* 90:101999
- Goel A, Sharma A, Mehra A (2018) Index tracking and enhanced indexing using mixed conditional value-at-risk. *J Comput Appl Math* 335:361–380
- Guastaroba G, Mansini R, Ogryczak W, Speranza M (2016) Linear programming models based on omega ratio for the enhanced index tracking problem. *Eur J Oper Res* 251(3):938–956
- Guastaroba G, Mansini R, Ogryczak v, Speranza M (2017) Enhanced index tracking with CVAR-based measures. In: Working Paper
- Koshizuka T, Konno H, Yamamoto R (2009) Index-plus-alpha tracking subject to correlation constraint. *Int J Optim: Theory Methods Appl* 1(2):215–224
- Kuosmanen T (2004) Efficient diversification according to stochastic dominance criteria. *Manag Sci* 50(10):1390–1406
- Lejeune M (2012) Game theoretical approach for reliable enhanced indexation. *Decision Anal* 9(2):146–155
- Lejeune M, Samatli-Pac G (2013) Construction of risk-averse enhanced index funds. *INFORMS J Comput* 25(4):701–719

- Li Q, Sun L, Bao L (2011) Enhanced index tracking based on multi-objective immune algorithm. *Expert Syst Appl* 38(5):6101–6106
- Luedtke J (2008) New formulations for optimization under stochastic dominance constraints. *SIAM J Optim* 19(3):1433–1450
- Mutunge P, Haugland D (2018) Minimizing the tracking error of cardinality constrained portfolios. *Comput Oper Res* 90:33–41
- Parvez A, Sudhir N (2005) Performance of enhanced index and quantitative equity funds. *Financ Rev* 40(4):459–479
- Post T, Kopa M (2017) Portfolio choice based on third-degree stochastic dominance. *Manag Sci* 63(10):3381–3392
- Roman D, Mitra G, Zverovich V (2013) Enhanced indexation based on second-order stochastic dominance. *Eur J Oper Res* 228(1):273–281
- Sant’Anna LR, Filomena TP, Guedes PC, Borenstein D (2017) Index tracking with controlled number of assets using a hybrid heuristic combining genetic algorithm and non-linear programming. *Ann Oper Res* 258(2):849–867
- Sharma A, Agrawal S, Mehra A (2017) Enhanced indexing for risk averse investors using relaxed second order stochastic dominance. *Optim Eng* 18(2):407–442
- Stoyan SJ, Kwon RH (2010) A two-stage stochastic mixed-integer programming approach to the index tracking problem. *Optim Eng* 11:247–275
- Valle C, Meade N, Beasley J (2014) Absolute return portfolios. *Omega* 45(1):20–41
- Weng Y-C, Wang R (2017) Do enhanced index funds truly have enhanced performance? Evidence from the chinese market. *Emerg Mark Finance Trade* 53(4):819–834
- Wu L-C, Chou S-C, Yang C-C, Ong C-S (2007) Enhanced index investing based on goal programming. *J Portf Manag* 33(3):49–56
- Xu F, Wang M, Dai Y-H, Xu D (2018) A sparse enhanced indexation model with chance and cardinality constraints. *J Glob Optim* 70(1):5–25

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.