**Submitted to *Operations Research***
**manuscript (Please, provide the manuscript number!)**



# A scalable algorithm for sparse and robust portfolios

Dimitris Bertsimas
Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA.
ORCID: 0000-0002-1985-1003
dbertsim@mit.edu

Ryan Cory-Wright
Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA,
ORCID: 0000-0002-4485-0619
ryancw@mit.edu

A central problem in financial economics concerns investing capital to maximize a portfolio's expected return while minimizing its variance, subject to an upper bound on the number of positions, minimum investment and linear inequality constraints, at scale. Existing approaches to this problem do not provide provably optimal portfolios for real-world problem sizes with more than 300 securities. In this paper, we present a cutting-plane method which solves problems with 1000s of securities to provable optimality, by exploiting a dual representation of the continuous Markowitz problem to obtain a closed form representation of the problem's subgradients. We improve the performance of the cutting-plane method in three different ways. First, we implement a local-search heuristic which applies our subgradient representation to obtain high-quality warm-starts. Second, we embed the local-search heuristic within the cutting-plane method to accelerate recovery of an optimal solution. Third, we exploit a correspondence between the convexified sparse Markowitz problem and a rotated second-order cone problem to obtain a tight lower bound which is sometimes exact. Finally, we establish that the cutting-plane method is 3—4 orders of magnitude more efficient than existing methods, and construct provably optimal sparsity-constrained frontiers for the S&P 500, Russell 1000, and Wilshire 5000.

*Key words*: Sparse Portfolio Optimization, Binary Convex Optimization, Robust Optimization.

---

*Subject classifications*: programming: integer; non-linear: quadratic; finance: portfolio
*Area of review*: Optimization

## 1. Introduction

Given an expected marginal return vector $\boldsymbol{\mu} \in \mathbb{R}^n$, estimated covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, uncertainty budget parameters $\sigma, \gamma > 0$, cardinality budget parameter $k \in \{2, ..., n-1\}$, linear constraint matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, and right-hand-side bounds $\boldsymbol{l}, \boldsymbol{u} \in \mathbb{R}^m$, investors determine an optimal allocation of capital between assets by solving the following mixed-integer quadratic optimization problem (MIQO, see Markowitz (1952)):

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{\sigma}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}+\frac{1}{2\gamma}||\boldsymbol{x}||_2^2-\boldsymbol{\mu}^\top\boldsymbol{x} \text{ s.t. } \boldsymbol{l}\leq\boldsymbol{A}\boldsymbol{x}\leq\boldsymbol{u},\ \boldsymbol{e}^\top\boldsymbol{x}=1,\ \boldsymbol{x}\geq\boldsymbol{0},\ ||\boldsymbol{x}||_0\leq k, \tag{1}$$

where $||\boldsymbol{x}||_0$ denotes the number of distinct positions held in the portfolio, which we refer to as the cardinality of the portfolio.

### 1.1. Background and literature review

Problem (1) is an NP-hard problem (Gao and Li (2013), Lemma 1) which was first studied by Bienstock (1996), who described a model for minimizing variance and expected disutility subject to constraints on investment. Additionally, Bienstock (1996) derived four valid inequalities and applied them to solve MIQOs via branch-and-cut.

Problem (1) was independently considered by Bertsimas et al. (1999), who proposed a similar model and applied it to generate high-quality portfolios for an investment management firm.

Problem (1) was subsequently studied by Chang et al. (2000), who demonstrated that simulated annealing, tabu search and genetic algorithms provide high-quality portfolios in an efficient manner. However, Chang et al. (2000) did not embed their heuristics within an optimization framework such as branch-and-bound, and consequently could not provide an optimality gap for their approach.

An algorithm for solving Problem (1) to provable optimality was subsequently developed by Shioda (2003) (see also Bertsimas and Shioda (2009)), who implemented Lemke's pivot method, embedded it within branch-and-bound, and demonstrated that it enjoys practical advantages over several of CPLEX's methods. However, they were unable to obtain informative lower bounds which intelligently discriminate between unexpanded nodes in a branch-and-bound tree. This limited their ability to solve Problem (1) to provable optimality with more than $n=50$ securities.

A perspective cut-based formulation for Problem (1) was subsequently developed by Frangioni and Gentile (2006), who demonstrated that perspective cuts provide substantially tighter relaxations than relaxing integrality on big-M constraints, and obtained a bound gap of $<1\%$ for problems with 200 securities.

A lifted branch-and-bound algorithm for solving Problem (1) was subsequently developed by Vielma et al. (2008), who demonstrated that algorithms which apply lifted polyhedral relaxations enjoy practical advantages over CPLEX. However, they were unable to reliably solve portfolio optimization instances with more than around $n=100$ securities to provable optimality.

A Markowitz portfolio optimization model with integrality and chance constraints was subsequently studied by Bonami and Lejeune (2009), who applied a nonlinear branch-and-bound algorithm to solve instances of Problem (1) with up to $n=200$ securities to provable optimality.

A partial solution to the lack of high-quality lower bounds was subsequently provided by Gao and Li (2013), who applied several Lagrangian relaxation procedures to obtain tighter lower bounds and embedded the Lagrangian relaxation procedure within a branch-and-bound framework (see also Cui et al. (2013), who developed a similar procedure using semidefinite optimization). However, they did not consider using cutting-planes to propagate the information obtained from solving sub-problems to other parts of the branch-and-bound tree, and consequently could not solve portfolio optimization instances with more than $n = 300$ securities.

More recently, Problem (1) was considered by Vielma et al. (2017), who derived a new extended formulation using a tower of variables reformulation, which has subsequently been implemented in the commercial solvers Gurobi (2018) and CPLEX (2018). However, this formulation cannot solve portfolio optimization instances with more than $n = 200$ securities to provable optimality.

We summarize the scalability of each approach in Table 1, and remind the reader that most real-world asset management problems comprise selecting $k$ securities from a universe of 500 to $5,000$ securities.

**Table 1** Largest instance of Problem (1) reliably solved, by approach.

| Reference | Solution method | Largest instance (no. securities) |
|---|---|---|
| Bertsimas and Shioda (2009) | Lemke pivot branch-and-bound | 50 |
| Frangioni and Gentile (2006) | Perspective cut | (1% bound gap) 200 |
| Vielma et al. (2008) | Lifted branch-and-bound | 100 |
| Bonami and Lejeune (2009) | Nonlinear branch-and-bound | 200 |
| Gao and Li (2013) | Lagrangian relaxation branch-and-bound | 300 |
| Cui et al. (2013) | Lagrangian relaxation branch-and-bound | 300 |
| Vielma et al. (2017) | Lifted branch-and-bound | 200 |

Due to both the difficulties encountered during the above attempts at solving the Markowitz model at scale, and Problem (1)'s NP-hardness, the greater finance community considers Problem (1) to be intractable and uses two approaches as surrogates. The first approach comprises applying heuristics to obtain high-quality solutions, as outlined by Chang et al. (2000). The second approach was independently applied to the Markowitz portfolio model by DeMiguel et al. (2009) and Brodie et al. (2009) and involves using a convex norm such as the $l_1$ norm as a surrogate for the $l_0$ norm. Applying this approach in the absence of nonnegativity constraints results in the following convex surrogate to Problem (1):

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \ \frac{\sigma}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} + \frac{1}{2\gamma}||\boldsymbol{x}||_2^2 - \boldsymbol{\mu}^\top \boldsymbol{x} + \rho||\boldsymbol{x}||_1 \text{ s.t. } \boldsymbol{l} \leq \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{u}, \ \boldsymbol{e}^\top \boldsymbol{x} = 1, \tag{2}$$

where $\rho$ is a parameter which the literature commonly claims promotes sparsity (see, e.g., Brodie et al. (2009)). Note that this approach cannot be applied in the presence of the nonnegativity constraint $\boldsymbol{x} \geq \boldsymbol{0}$, since it implies that $\rho||\boldsymbol{x}||_1 = \rho$, wherever $\boldsymbol{x}$ is feasible.

The convex relaxation encapsulated in Problem (2) provides an optimal solution to Problem (1) under some strong assumptions which typically cannot be verified without performing a procedure equivalent to solving Problem (1) (see Wainwright (2009) for conditions under which the $l_1$ norm recovers an optimal solution in sparse regression). Moreover, the parameter $\rho$ promotes robustness, rather than sparsity. We formalize this observation in the following theorem:

THEOREM 1. *The following two optimization problems are equivalent:*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \max_{||\Delta\boldsymbol{\mu}||_\infty\le\rho} \quad \frac{\sigma}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}+\frac{1}{2\gamma}||\boldsymbol{x}||_2^2-(\boldsymbol{\mu}+\Delta\boldsymbol{\mu})^\top\boldsymbol{x} \quad s.t. \quad \boldsymbol{l}\le\boldsymbol{A}\boldsymbol{x}\le\boldsymbol{u},\ \boldsymbol{e}^\top\boldsymbol{x}=1,$$
$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad \frac{\sigma}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}+\frac{1}{2\gamma}||\boldsymbol{x}||_2^2-\boldsymbol{\mu}^\top\boldsymbol{x}+\rho||\boldsymbol{x}||_1 \quad s.t. \quad \boldsymbol{l}\le\boldsymbol{A}\boldsymbol{x}\le\boldsymbol{u},\ \boldsymbol{e}^\top\boldsymbol{x}=1.$$

*Proof of Theorem 1* See Ben-Tal et al. (2009). □

Theorem 1 indicates that the strategy of using a $l_1$ norm as a surrogate for a $l_0$ norm leads to portfolios which are robust to data errors in the expected return vector $\boldsymbol{\mu}$, thus improving their out-of-sample performance. However, without strong assumptions akin to those made in lasso regression (see, e.g., Wainwright (2009)), optimal solutions to Problem (2) are unlikely to solve Problem (1).

Traditionally, exact approaches to Problem (1) comprise using big-M constraints to determine an optimal subset of securities (see, e.g., Bertsimas and Shioda (2009)). The following MIQO illustrates this approach:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\ \boldsymbol{s}\in S_n^k} \quad \frac{\sigma}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}-\boldsymbol{\mu}^\top\boldsymbol{x}+\frac{1}{2\gamma}||\boldsymbol{x}||_2^2 \quad \text{s.t.} \quad \boldsymbol{l}\le\boldsymbol{A}\boldsymbol{x}\le\boldsymbol{u},\ \boldsymbol{e}^\top\boldsymbol{x}=1,\ \boldsymbol{0}\le\boldsymbol{x}\le\boldsymbol{s}, \tag{3}$$

where the vector of binary variables $\boldsymbol{s}$ imposes the sparsity constraint, we write $\boldsymbol{x}\le\boldsymbol{s}$ rather than $\boldsymbol{x}\le M\boldsymbol{s}$ because the unit simplex constraint implies $M=1$ and $S_n^k:=\{\boldsymbol{s}\in\{0,1\}^n:\boldsymbol{e}^\top\boldsymbol{s}\le k\}$ is the set of $n$-dimensional binary vectors with at most $k$ non-zero components.

Problem (3) is mathematically equivalent to Problem (1). However, as identified by Bertsimas and Van Parys (2016), imposing sparsity via the constraint $\boldsymbol{x}\le\boldsymbol{s}$ provides lower bounds which are too weak to prune most nodes in branch-and-bound trees which arise from Problem (3).

Recently, a new approach was identified by Bertsimas and Van Parys (2016) in the context of sparse regression, which avoids the computational difficulties which occur in solving MIQOs via a big-M approach. This approach comprises reformulating the problem as a saddle-point optimization problem with an outer linear integer optimization problem and an inner dual quadratic optimization problem, and designing a cutting-plane method which exploits the saddle-point structure. This paper generalizes the new approach to a broader class of MIQOs and applies it to solve large-scale instances of Problem (1) to provable optimality.

### 1.2. Contributions and structure

The key contribution of the paper is a new algorithm that allows us to solve large scale sparse portfolio optimization problems with up to $5,000$ securities to provable optimality.

The structure of the paper is as follows:

- In Section 2, we provide some structural properties of Problem (1) including its equivalence to a robust optimization problem and worst-case performance guarantees. We also describe the relationship between the hyperparameters and optimal solutions to Problem (1).
- In Section 3, we provide a dual perspective on Problem (1)'s continuous relaxation, and demonstrate that solving its inner dual yields subgradients with respect to the positions held.
- In Section 4, we design a cutting-plane method which solves Problem (1) to provable optimality.
- In Section 5, we improve the cutting-plane method in three ways. First, we describe a local-search heuristic which provides high-quality warm-starts. Second, we embed the local-search heuristic within the cutting-plane method to accelerate recovery of an optimal solution. Third, we observe that Problem (1)'s continuous relaxation corresponds to a Second Order Cone Problem, which provides a lower bound which is often near-exact and sometimes exact, and demonstrate that exactness of the SOCP bound is sufficient to recover an optimal solution to Problem (1) apriori.
- In Section 6, we apply the cutting-plane method to the problems described in Chang et al. (2000), and three larger scale problems: the S&P 500, Russell 1000, and Wilshire 5000. We also explore Problem (1)'s sensitivity to its hyperparameters, and establish empirically that optimal solutions tend to be stable for reasonable hyperparameter choices.

#### Notation

We use nonbold face characters to denote scalars, lowercase bold faced characters such as $\boldsymbol{x} \in \mathbb{R}^n$ to denote vectors, and uppercase bold faced characters such as $\boldsymbol{X} \in \mathbb{R}^{n \times r}$ to denote matrices. We let $\mathbf{e}$ denote a vector of all 1's, $\mathbf{0}$ denote a vector of all 0's, and $\mathbb{I}$ denote an identity matrix, with dimension implied by the context. We let $S_n^k$ denote the set of all $n$-dimensional binary vectors with $k$ entries equal to 1, that is,

$$S_n^k := \{\boldsymbol{s} \in \{0,1\}^n : \boldsymbol{e}^\top \boldsymbol{s} \leq k\}.$$

We denote the number of non-zero elements in a vector by the support norm $l_0$. Finally, we let $\mathbb{R}_+^n$ denote the $n$-dimensional nonnegative orthant.

## 2. Motivation and interpretation

In this section, we provide some properties of Problem (1), including its probabilistic performance guarantees and the interpretation of the robustness parameters.

### 2.1. Robust portfolios

The literature commonly justifies investing capital according to optimal solutions of Problem (1) by interpreting (a) variance as a measure of risk which the coefficient $\sigma$ prices against the expected payoff $\boldsymbol{\mu}^\top \boldsymbol{x}$, and (b) the term $||\boldsymbol{x}||_2^2$ as a regularization of the amount invested in each security, to obtain superior out-of-sample performance. However, Problem (1) is more naturally motivated by its equivalence to the problem of maximizing the worst-case expected return for a given uncertainty budget. We formalize this observation in the following theorem:

THEOREM 2. *The following two optimization problems are equivalent:*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \max_{\left|\left|(\boldsymbol{\Sigma}+\frac{1}{\gamma}\mathbb{I})^{-1/2}\Delta\boldsymbol{\mu}\right|\right|\leq \frac{\sigma}{2}} -(\boldsymbol{\mu}+\Delta\boldsymbol{\mu})^\top \boldsymbol{x} \text{ s.t. } \boldsymbol{l}\leq \boldsymbol{A}\boldsymbol{x}\leq \boldsymbol{u},\ \boldsymbol{e}^\top\boldsymbol{x}=1,\ \boldsymbol{x}\geq \boldsymbol{0},\ ||\boldsymbol{x}||_0\leq k, \tag{4a}$$

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{\sigma}{2}\left|\left|\left(\boldsymbol{\Sigma}+\frac{1}{\gamma}\mathbb{I}\right)^{\frac{1}{2}}\boldsymbol{x}\right|\right| -\boldsymbol{\mu}^\top \boldsymbol{x} \text{ s.t. } \boldsymbol{l}\leq \boldsymbol{A}\boldsymbol{x}\leq \boldsymbol{u},\ \boldsymbol{e}^\top\boldsymbol{x}=1,\ \boldsymbol{x}\geq \boldsymbol{0},\ ||\boldsymbol{x}||_0\leq k. \tag{4b}$$

*Proof of Theorem 2* This result follows from applying Lagrangian duality and manipulating Problem (4a)'s KKT conditions, as outlined in Ben-Tal et al. (2009). $\square$

Theorem 2 demonstrates that the Markowitz portfolio model is a robust optimization problem. This result is supported by a large body of literature on probabilistic guarantees. For instance, Bertsimas and Popescu (2005) have shown that under any distribution of payoffs with first moment $\boldsymbol{\mu}$ and second central moment $\boldsymbol{\Sigma}$, solutions to Problem (1) provide a payoff of at least $\boldsymbol{\mu}^\top x - \sqrt{\frac{1-\epsilon}{\epsilon}}\sqrt{\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}}$ with probability $1-\epsilon$, and moreover the expected payoff in the other $\epsilon$ % of scenarios is never worse than $\boldsymbol{\mu}^\top x - \sqrt{\frac{1-\epsilon}{\epsilon}}\sqrt{\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}}$. We note that terms in Problem (4b)'s objective are different from their counterparts in Problem (1) by a factor of a square-root, meaning we must solve Problem (1) a small number of times to solve Problem (4b) exactly. However, this is not an issue in practice, where we choose Problem (1)'s hyperparameters to maximize the out-of-sample risk-adjusted expected profit, and obtain an ex-post robustness guarantee from Theorem 2.

The interpretation of each robustness parameter is as follows:

- $\gamma$ is called the Tikhonov regularization parameter Tikhonov (1963), and typically takes values between $\frac{1}{n}$ and $\frac{1}{\sqrt{n}}$. As $\gamma \mapsto 0$, the regularized covariance matrix $\frac{1}{\gamma}\mathbb{I}+\boldsymbol{\Sigma}$ becomes diagonally-dominant and (provided $\boldsymbol{l}\leq\boldsymbol{A}\boldsymbol{x}\leq\boldsymbol{u}$ is non-binding) optimal solutions to Problem (1) correspond to cardinality-constrained versions of the $\frac{1}{N}$ portfolio optimization strategy studied by DeMiguel et al. (2007) and Pflug et al. (2012), wherein $\frac{1}{k}$ of the total investment is allocated to each one of the $k$ securities with the highest expected return. Alternatively, setting $\gamma\mapsto\infty$ recovers the classical Markowitz model from Problem (1), meaning our approach has a Bayesian interpretation wherein we initially embrace the $\frac{1}{k}$ investment strategy and update this belief as we receive more data.

- $\sigma$ is called the Arrow-Pratt risk-aversion coefficient (see Arrow (1965), Pratt (1975)), and typically takes values between 1 and 10. Setting $\sigma \mapsto \infty$ corresponds to distrusting the estimated return vector $\boldsymbol{\mu}$ and instead picking the $k$ securities which provide the minimum variance portfolio. This might be a valid strategy in situations with high parameter uncertainty, as typically estimates of the covariance matrix $\boldsymbol{\Sigma}$ are more stable than estimates of the return vector $\boldsymbol{\mu}$. Alternatively, setting $\sigma \mapsto 0$ constitutes maintaining a high degree of confidence that the return vector is not contaminated with errors. In this case, an optimal solution to Problem (1) comprises picking the security with the highest expected return (provided this strategy satisfies the constraint $\boldsymbol{l} \leq \boldsymbol{Ax} \leq \boldsymbol{u}$). Finally, a choice of $\sigma$ between these extremes constitutes a trade-off between maximizing a portfolio's in-sample expected return and mitigating the impact of estimation errors on the portfolio's out-of-sample performance.

### 2.2. Sparse portfolios

A second issue in portfolio optimization is sparsity, i.e., imposing the constraint that $||x||_0 \leq k$. Portfolio managers require sparse portfolios for the following three reasons:

1. Managers incur transaction costs whenever they alter their position in a security, and this cost increases with the number of positions held.
2. Managers incur monitoring costs for each non-zero positions held.
3. Investors believe that fund managers make two decisions: which positions to hold, and the proportion of funds allocated to each position. Consequently, investing in the entire market is perceived as equivalent to performing index-tracking while charging active management fees.

Uniting these observations with the preceding discussion on robustness leads to Problem (1).

### 2.3. Equivalence between portfolio selection and regression

As noted by authors including Brodie et al. (2009) and Olivares-Nadal and DeMiguel (2018), the covariance matrix $\boldsymbol{\Sigma}$ can be decomposed as $\boldsymbol{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}$, where $\boldsymbol{\Sigma}$ is possibly rank deficient. Consequently, by scaling $\boldsymbol{\Sigma} \hookleftarrow \sigma \boldsymbol{\Sigma}$ and letting:

$$\boldsymbol{y} := (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{\mu}, \tag{5}$$

$$\boldsymbol{d} := (\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X} - \mathbb{I})\boldsymbol{\mu}, \tag{6}$$

be the projection of the return vector $\boldsymbol{\mu}$ onto the span and nullspace of $\boldsymbol{X}$, we rewrite Problem (1) as the following constrained regression problem, where we add the constant term $\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{y}$ without loss of generality:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad \frac{1}{2\gamma}||\boldsymbol{x}||_2^2 + \frac{1}{2}\ ||\boldsymbol{Xx} - \boldsymbol{y}||_2^2 + \boldsymbol{d}^\top \boldsymbol{x} \text{ s.t} \quad \boldsymbol{l} \leq \boldsymbol{Ax} \leq \boldsymbol{u},\ \boldsymbol{e}^\top \boldsymbol{x} = 1,\ \boldsymbol{x} \geq \boldsymbol{0},\ ||\boldsymbol{x}||_0 \leq k. \tag{7}$$

Problem (7) provides a link between Problem (1) and the class of sparse regression problems studied by Bertsimas and Van Parys (2016). This connection is notable, because Bertsimas and Van Parys (2016) describes a cutting-plane method which solves sparse regression problems with 100,000s of variables to optimality in seconds. Motivated by this observation, we proceed to derive an analogous cutting-plane method which solves a more general class of MIQOs.

## 3. A dual perspective on Markowitz portfolio optimization

In this section, we momentarily ignore the discreteness introduced by the cardinality constraint and instead focus on the following continuous optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad \frac{1}{2\gamma}||\boldsymbol{x}||_2^2 + \frac{\sigma}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x} - \boldsymbol{\mu}^\top\boldsymbol{x} \text{ s.t. } \boldsymbol{l} \le \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{u},\ \boldsymbol{e}^\top\boldsymbol{x} = 1,\ \boldsymbol{x} \ge \boldsymbol{0}, \tag{8}$$

which, by Section 2.3, is equivalent to the following problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad \frac{1}{2\gamma}||\boldsymbol{x}||_2^2 + \frac{1}{2}\ ||\boldsymbol{X}\boldsymbol{x} - \boldsymbol{y}||_2^2 + \boldsymbol{d}^\top\boldsymbol{x} \text{ s.t } \boldsymbol{l} \le \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{u},\ \boldsymbol{e}^\top\boldsymbol{x} = 1,\ \boldsymbol{x} \ge \boldsymbol{0}, \tag{9}$$

with a view to show that Problem (9) has the following dual problem:

$$\begin{aligned}
\max_{\substack{\boldsymbol{\pi}\in\mathbb{R}^n_+,\ \boldsymbol{\alpha}\in\mathbb{R}^r,\ \boldsymbol{w}\in\mathbb{R}^n,\\ \boldsymbol{\beta}_l,\ \boldsymbol{\beta}_u\in\mathbb{R}^m_+,\ \lambda\in\mathbb{R}}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} - \frac{\gamma}{2}\boldsymbol{w}^\top\boldsymbol{w} + \boldsymbol{y}^\top\boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top\boldsymbol{l} - \boldsymbol{\beta}_u^\top\boldsymbol{u} + \lambda \\
\text{s.t.} \quad & \boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda\boldsymbol{e} + \boldsymbol{A}^\top(\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.
\end{aligned} \tag{10}$$

Problem (9) is equivalent to the following optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{x}\in\mathbb{R}^n,\ \boldsymbol{r}\in\mathbb{R}^r} \quad & \frac{1}{2\gamma}||\boldsymbol{x}||_2^2 + \frac{1}{2}\ ||\boldsymbol{r}||_2^2 + \boldsymbol{d}^\top\boldsymbol{x} \\
\text{s.t} \quad & \boldsymbol{X}\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{r}, && [\boldsymbol{\alpha}], \\
& \boldsymbol{A}\boldsymbol{x} \ge \boldsymbol{l}, && [\boldsymbol{\beta}_l], \\
& \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{u}, && [\boldsymbol{\beta}_u], \\
& \boldsymbol{e}^\top\boldsymbol{x} = 1, && [\lambda], \\
& \boldsymbol{x} \ge \boldsymbol{0}, && [\boldsymbol{\pi}],
\end{aligned} \tag{11}$$

where each primal constraint is matched with a vector of dual variables in square brackets.

We assume throughout this paper that Problem (11) satisfies a constraint qualification, and therefore is equivalent to solving its KKT conditions. To avoid imposing a choice of constraint qualification, we formalize this in the following general assumption:

Assumption 1. Problem (11) satisfies a constraint qualification.

A constraint qualification which satisfies Assumption 1 is Slater's Constraint Qualification (see Appendix D of Ben-Tal and Nemirovski (2001)), which holds if the following set is non-empty:

$$\{\boldsymbol{x} \in \mathbb{R}^n |\ l_j < \boldsymbol{a}_j^\top\boldsymbol{x} < u_j\ \forall j \text{ s.t. } l_j < u_j,\ \boldsymbol{a}_j^\top\boldsymbol{x} = u_j\ \forall j \text{ s.t. } l_j = u_j\}.$$

If $\boldsymbol{l} \leq \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{u}$ enforces the real-world constraints outlined by Bertsimas et al. (1999) requiring the investment in each market sector to be near a benchmark index, then Assumption 1 is equivalent to primal feasibility. Alternatively, with the chance constraints studied by Bonami and Lejeune (2009) requiring a portfolio to exceed some benchmark return with high probability, Assumption 1 can be verified by solving a separation problem (see Ben-Tal and Nemirovski (2001)).

After assuming that a constraint qualification holds, we have the following theorem:

THEOREM 3. *Suppose Assumption 1 holds. Then, Problem (11) has the following dual problem:*

$$\begin{aligned}
\max_{\substack{\boldsymbol{\pi}\in\mathbb{R}^n_+,\ \boldsymbol{\alpha}\in\mathbb{R}^r,\ \boldsymbol{w}\in\mathbb{R}^n,\\ \boldsymbol{\beta}_l,\ \boldsymbol{\beta}_u\in\mathbb{R}^m_+,\ \lambda\in\mathbb{R}}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} - \frac{\gamma}{2}\boldsymbol{w}^\top\boldsymbol{w} + \boldsymbol{y}^\top\boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top\boldsymbol{l} - \boldsymbol{\beta}_u^\top\boldsymbol{u} + \lambda \\
\text{s.t.} \quad & \boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda\boldsymbol{e} + \boldsymbol{A}^\top(\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.
\end{aligned}$$

*Proof of Theorem 3* By Assumption 1, we can solve Problem (11) by minimizing a Lagrangian, which is given by the following expression:

$$\mathcal{L} \quad = \frac{1}{2\gamma}\boldsymbol{x}^\top\boldsymbol{x} + \frac{1}{2}\boldsymbol{r}^\top\boldsymbol{r} + \boldsymbol{d}^\top\boldsymbol{x} - \boldsymbol{\alpha}^\top(\boldsymbol{X}\boldsymbol{x} - \boldsymbol{y} - r) - \boldsymbol{\pi}^\top\boldsymbol{x} - \lambda(\boldsymbol{e}^\top\boldsymbol{x} - 1) - \boldsymbol{\beta}_l^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{l}) + \boldsymbol{\beta}_u^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{u}).$$

Moreover, the Lagrangian $\mathcal{L}$ is equivalent to solving the following KKT conditions:

$$\begin{aligned}
\nabla_{\boldsymbol{x}}\mathcal{L} = \boldsymbol{0} &\implies \frac{1}{\gamma}\boldsymbol{x} + \boldsymbol{d} - \boldsymbol{X}^\top\boldsymbol{\alpha} - \boldsymbol{\pi} - \lambda\boldsymbol{e} - \boldsymbol{A}^\top(\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) = 0, \\
&\implies \boldsymbol{x} = \gamma(\boldsymbol{X}^\top\boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda\boldsymbol{e} + \boldsymbol{A}^\top(\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}), \\
\nabla_{\boldsymbol{r}}\mathcal{L} = \boldsymbol{0} &\implies \boldsymbol{r} - \boldsymbol{\alpha} = 0 \implies \boldsymbol{r} = \boldsymbol{\alpha}.
\end{aligned}$$

Substituting the above expressions for $\boldsymbol{x}, \boldsymbol{r}$ into $\mathcal{L}$ and rearranging to maximize provides the stated dual problem, where we impose the constraints $\boldsymbol{\beta}_l, \boldsymbol{\beta}_u \geq \boldsymbol{0}, \boldsymbol{\pi} \geq \boldsymbol{0}$ because these Lagrange multipliers are coupled with inequality constraints, and we introduce the dummy variable $\boldsymbol{w}$ for brevity. ☐

The derivation of Problem (10) reveals that each optimal allocation of funds $\boldsymbol{x}^*$ satisfies the following relationship for some set of optimal dual variables:

$$\boldsymbol{x}^* = \gamma(\boldsymbol{X}^\top\boldsymbol{\alpha}^* + \boldsymbol{\pi}^* + \boldsymbol{\lambda}^*\boldsymbol{e} + \boldsymbol{A}^\top(\boldsymbol{\beta}_l^* - \boldsymbol{\beta}_u^*) - \boldsymbol{d}). \tag{12}$$

Equation (12) permits recovery of an optimal primal solution from a set of optimal dual variables.

In the next section, we proceed to reintroduce the sparsity constraint and develop a cutting-plane method which solves Problem (1) to provable optimality.

## 4. A cutting-plane method

The starting point of this section is the observation that we can leverage the previous section's analysis to provide a dual perspective on Problem (1), by rewriting it in the following form:

$$\min_{\boldsymbol{s}\in S_k^n} \Big[f(\boldsymbol{s})\Big], \tag{13}$$

where we define:

$$f(\boldsymbol{s}) := \min_{\boldsymbol{x}\in\mathbb{R}^{\sum_i s_i}} \quad \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{x} + \frac{1}{2}||\boldsymbol{X_s x} - \boldsymbol{y}||_2^2 + \boldsymbol{d_s}^\top\boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{l} \le \boldsymbol{A_s x} \le \boldsymbol{u},\ \boldsymbol{e_s}^\top\boldsymbol{x} = 1,\ \boldsymbol{x} \ge \boldsymbol{0}, \tag{14}$$

$\boldsymbol{X_s}$ is a matrix which contains the columns of $\boldsymbol{X}$ selected by the vector $\boldsymbol{s}$, and $\boldsymbol{A_s}, \boldsymbol{d}_s, \boldsymbol{e}_s$ are similar.

We combine this observation with Theorem 3 to yield the following theorem:

THEOREM 4. *Suppose Assumption 1 holds. Then, Problem (13)'s continuous relaxation is equivalent to the following saddle point problem:*

$$\begin{aligned}\min_{\boldsymbol{s}\in \mathrm{Conv}(S_k^n)} \max_{\substack{\boldsymbol{\pi}\in\mathbb{R}^n_+,\ \boldsymbol{\alpha}\in\mathbb{R}^r,\ \boldsymbol{w}\in\mathbb{R}^n,\\ \boldsymbol{\beta}_l,\ \boldsymbol{\beta}_u\in\mathbb{R}^m_+,\ \lambda\in\mathbb{R}}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} - \frac{\gamma}{2}\sum_i s_i w_i^2 + \boldsymbol{y}^\top\boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top\boldsymbol{l} - \boldsymbol{\beta}_u^\top\boldsymbol{u} + \lambda \\ \textit{s.t.} \quad & \boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda\boldsymbol{e} + \boldsymbol{A}^\top(\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d}.\end{aligned} \tag{15}$$

*Proof of Theorem 4* By combining Theorem 3 with the Sion-Kakutani minimax theorem (see Ben-Tal and Nemirovski (2001)), for any feasible subset $\boldsymbol{s}$, we have that:

$$\begin{aligned}f(\boldsymbol{s}) = \max_{\substack{\boldsymbol{\pi_s}\in\mathbb{R}^{\hat{k}}_+,\ \boldsymbol{\alpha}\in\mathbb{R}^r,\ \boldsymbol{w_s}\in\mathbb{R}^{\hat{k}},\\ \boldsymbol{\beta}_l,\ \boldsymbol{\beta}_u\in\mathbb{R}^m_+,\ \lambda\in\mathbb{R}}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} - \frac{\gamma}{2}\boldsymbol{w_s}^\top\boldsymbol{w_s} + \boldsymbol{y}^\top\boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top\boldsymbol{l} - \boldsymbol{\beta}_u^\top\boldsymbol{u} + \lambda \\ \text{s.t.} \quad & \boldsymbol{w_s} = \boldsymbol{X_s}^\top\boldsymbol{\alpha} + \boldsymbol{\pi_s} + \lambda\boldsymbol{e_s} + \boldsymbol{A_s}^\top(\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d_s},\end{aligned}$$

where $\hat{k} := \sum_i s_i$.

Moreover, this is equivalent to solving the following optimization problem:

$$\begin{aligned}f(\boldsymbol{s}) = \max_{\substack{\boldsymbol{\pi}\in\mathbb{R}^n_+,\ \boldsymbol{\alpha}\in\mathbb{R}^r,\ \boldsymbol{w}\in\mathbb{R}^n,\\ \boldsymbol{\beta}_l,\ \boldsymbol{\beta}_u\in\mathbb{R}^m_+,\ \lambda\in\mathbb{R}}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} - \frac{\gamma}{2}\sum_i s_i w_i^2 + \boldsymbol{y}^\top\boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top\boldsymbol{l} - \boldsymbol{\beta}_u^\top\boldsymbol{u} + \lambda \\ \text{s.t.} \quad & \boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda\boldsymbol{e} + \boldsymbol{A}^\top(\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d},\end{aligned}$$

where we strengthen the formulation by associating $s_i$, rather than $s_i^2$, with $w_i^2$.

Minimizing $\boldsymbol{s}$ over $S_n^k$ then yields the result. □

It should be noted that for a fixed $\boldsymbol{s}$, Problem (15) has a unique optimal dual solution, since the inner optimization problem is strictly concave. However, there may be multiple optimal subsets $\boldsymbol{s}$. This idea of conditional uniqueness is implicit in the following corollaries to Theorem 4, which provide efficient objective function evaluations and subgradients:

COROLLARY 1. *Let $\boldsymbol{\pi}^*$, $\boldsymbol{\beta}_l^*$, $\boldsymbol{\beta}_u^*$, $\lambda^*$, $\boldsymbol{\alpha}^*$ be the optimal dual multipliers for a given subset of securities $\hat{s}$. Then, the objective function $f(\hat{s})$ is given by the following expression:*

$$f(\hat{\boldsymbol{s}}) = -\frac{\gamma}{2}||\boldsymbol{X}^\top \boldsymbol{\alpha}^* + \boldsymbol{\pi}^* + \lambda^* \boldsymbol{e} + \boldsymbol{A}^\top(\boldsymbol{\beta}_l^* - \boldsymbol{\beta}_u^*) - \boldsymbol{d}||_2^2 - \frac{1}{2}||\boldsymbol{\alpha}^*||_2^2 + \boldsymbol{y}^\top \boldsymbol{\alpha}^* + \lambda^* + \boldsymbol{l}^\top \boldsymbol{\beta}_l^* - \boldsymbol{u}^\top \boldsymbol{\beta}_u^*. \tag{16}$$

COROLLARY 2. *Let $\boldsymbol{\pi}^*$, $\boldsymbol{\beta}_l^*$, $\boldsymbol{\beta}_u^*$, $\lambda^*$, $\boldsymbol{\alpha}^*$ be the optimal dual multipliers for a particular subset of securities $\hat{s}$. Then, valid subgradients with respect to each security $i$ are given by the following expression:*

$$\frac{\partial f(\boldsymbol{s})}{\partial s_i} = -\frac{\gamma}{2}(\boldsymbol{X}_i^\top \boldsymbol{\alpha}^* + \boldsymbol{A}_i^\top(\boldsymbol{\beta}_l^* - \boldsymbol{\beta}_u^*) + \lambda^* + \pi_i^* - d_i)^2. \tag{17}$$

Before formalizing the above corollaries in an algorithmic procedure, there is one further matter to attend to. For computational efficiency purposes, we require that our cutting-plane method comprises solving a sequence of $k \times k$-sized subproblems, while $\boldsymbol{\pi} \in \mathbb{R}_+^n$. Fortunately, the optimal choice of dual variable is $\pi_i^* = 0$ whenever $\hat{s}_i = 0$, because we require that the KKT condition $\boldsymbol{0} \leq \boldsymbol{x} \perp \boldsymbol{\pi} \geq \boldsymbol{0}$ remains satisfied when $s_i > 0$ for the subgradient to be valid.

Corollaries 1—2 provide the first two terms in the Taylor series expansion of Problem (1)'s continuous relaxation. As some subsets of securities $\hat{\boldsymbol{s}}$ may not provide feasible allocations of funds between securities, which can be detected by unbounded dual problems, combining Corollaries 1 and 2 with feasibility cuts which ban infeasible solutions $\hat{\boldsymbol{s}}$ by imposing the following constraint:

$$\sum_i \hat{s}_i(1 - s_i) + \sum_i (1 - \hat{s}_i) s_i \geq 1 \tag{18}$$

yields a branch-and-cut variant of Kelley (1960)'s cutting-plane method, which solves the following epigraph version of Problem (15):

$$\min_{\theta \in \mathbb{R},\ \boldsymbol{s} \in S_n^k} \quad \theta \quad \text{s.t.} \quad \theta \geq f(\hat{\boldsymbol{s}}) + \nabla f(\hat{\boldsymbol{s}})^\top(\hat{\boldsymbol{s}} - \boldsymbol{s}),\ \forall \hat{\boldsymbol{s}} \in S_n^k:\ \hat{\boldsymbol{s}} \text{ is feasible}. \tag{19}$$

We outline the corresponding cutting-plane method in Algorithm 1. Note that we add the valid inequalities via lazy callbacks to maintain a single tree of partial solutions throughout the entire branch-and-cut process, and avoid the cost otherwise incurred in rebuilding the tree whenever a cut is added to Problem (1).

**Algorithm 1** Kelley (1960)'s cutting-plane method for Problem (1).

**procedure** Cutting-plane method
**input:** $\boldsymbol{X} \in \mathbb{R}^{r\times n},\ \boldsymbol{y} \in \mathbb{R}^r,\ \boldsymbol{d} \in \mathbb{R}^n,\ \boldsymbol{A} \in \mathbb{R}^{m\times n},\ \boldsymbol{l} \in \mathbb{R}^m,\ \boldsymbol{u} \in \mathbb{R}^m,\ \gamma \in \mathbb{R}$ and $k \in \{2, ..., n-1\}$
**output:** $\boldsymbol{s}^* \in S_n^k$ and $\boldsymbol{x}^* \in \mathbb{R}^n$
  **while** $\theta_t < f(\boldsymbol{s}_t)$ **do**
    $\boldsymbol{\alpha}_t^*, \boldsymbol{\pi}_t^*, \lambda_t^*, \boldsymbol{\beta}_{l,t}^*, \boldsymbol{\beta}_{u,t}^* \leftarrow \text{OptDuals}(\boldsymbol{A}_s,\ \boldsymbol{X}_s,\ \boldsymbol{d}_s)$ ▷ Solve Problem (10)
    **if** Dual Unbounded **then** ▷ Dual always feasible.
      Add feasibility cut $\sum_i \hat{s}_i(1-s_i) + \sum_i (1-\hat{s}_i)s_i \geq 1$
    **else**
      $f(\boldsymbol{s}_t) \leftarrow -\frac{\gamma}{2}||\boldsymbol{X}_s^\top \boldsymbol{\alpha}_t^* + \lambda_t^* \boldsymbol{e}_s + \boldsymbol{\pi}_t^* + \boldsymbol{A}_s^\top(\boldsymbol{\beta}_{l,t}^* - \boldsymbol{\beta}_{u,t}^*) - \boldsymbol{d}_s||_2^2$
      $\qquad -\frac{1}{2}||\boldsymbol{\alpha}_t^*||_2^2 + \boldsymbol{y}^\top \boldsymbol{\alpha}_t^* + \boldsymbol{l}^\top \boldsymbol{\beta}_{l,t}^* - \boldsymbol{u}^\top \boldsymbol{\beta}_{u,t}^* + \lambda_t^*$
      **for** $i \in \{1, ..., n\}$ **do**
        $\nabla f(\boldsymbol{s}_t)_i \leftarrow -\frac{\gamma}{2}(\boldsymbol{X}_{\boldsymbol{i}}^\top \boldsymbol{\alpha}_t^* + \boldsymbol{A}_i^\top(\boldsymbol{\beta}_{l,t}^* - \boldsymbol{\beta}_{u,t}^*) + \lambda_t^* + \pi_{i,t}^* - d_i)^2$
    $\boldsymbol{s}_{t+1}, \theta_{t+1} \leftarrow \arg\min_{\boldsymbol{s},\theta} \left\{\theta \text{ s.t. } \boldsymbol{s} \in S_n^k,\ \theta \geq f(\boldsymbol{s}_j) + \nabla f(\boldsymbol{s}_j)^\top(\boldsymbol{s}-\boldsymbol{s}_j), \forall j \in \{1,...,t\},\ \text{Feasibility cuts}\right\}$
    $t \leftarrow t+1$
  $\boldsymbol{s}^* \leftarrow \boldsymbol{s}_t,\ \boldsymbol{x}^* \leftarrow \boldsymbol{0},\ \boldsymbol{x}_s^* \leftarrow \gamma(\boldsymbol{X}_s^\top \boldsymbol{\alpha}_t^* + \boldsymbol{A}_s^\top(\boldsymbol{\beta}_{l,t}^* - \boldsymbol{\beta}_{u,t}^*) + \lambda_t^* \boldsymbol{e}_s + \boldsymbol{\pi}_t^* - \boldsymbol{d}_s)$

Algorithm 1 assumes that investors restrict their investment to a subset of at most $k$ securities. However, investors might prefer explicitly modelling the cost of monitoring each position. In this case, replacing the master problem constraint $\boldsymbol{e}^\top \boldsymbol{s} \leq k$ with the objective term $\sum_i \delta_i s_i$ yields the following saddle-point problem, where $\delta_i$ is the possibly distinct cost of monitoring each position:

$$\begin{aligned} \min_{\boldsymbol{s} \in \text{Conv}(S_n^n)} \quad & \sum_i \delta_i s_i + \Big[ \max_{\substack{\boldsymbol{\pi} \in \mathbb{R}_+^n,\ \boldsymbol{\alpha} \in \mathbb{R}^r,\ \boldsymbol{w} \in \mathbb{R}^n, \\ \boldsymbol{\beta}_l,\ \boldsymbol{\beta}_u \in \mathbb{R}_+^m,\ \lambda \in \mathbb{R}}} -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{\gamma}{2}\sum_i s_i w_i^2 + \boldsymbol{y}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}_l^\top \boldsymbol{l} - \boldsymbol{\beta}_u^\top \boldsymbol{u} + \lambda \\ \text{s.t.} \quad & \qquad \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\pi} + \lambda \boldsymbol{e} + \boldsymbol{A}^\top(\boldsymbol{\beta}_l - \boldsymbol{\beta}_u) - \boldsymbol{d} \Big]. \end{aligned} \tag{20}$$

It should be noted that Problems (15) and (20) are not equivalent, due to $\boldsymbol{s}$'s integrality.

The following theorem provides a proof of convergence for Algorithm 1:

Theorem 5. *Every limit point $\{\boldsymbol{x}^k, \boldsymbol{s}^k\}$ of a sequence generated by Algorithm 1 is an optimal solution to Problem (1). Moreover, an optimal solution is obtained in a finite number of iterations.*

*Proof of Theorem 5* See Theorem 2 of Fletcher and Leyffer (1994). □

We remind the reader that the optimal limit point may not be unique. However, the allocation of funds $\boldsymbol{x}^*$ is conditionally unique given $\boldsymbol{s}^*$, since the inner minimization problem is strictly convex.

Unfortunately, while Theorem 5 reassures us that Algorithm 1 terminates in a finite number of iterations, the polyhedron defined by $\text{Conv}(S_n^k)$ has an exponential number of extreme points. Therefore, in the worst case, Algorithm 1 generates an exponential number of cuts. Indeed, the worst-case performance of cutting-plane methods is arbitrarily bad in general (Nesterov (2013), Example 3.3.2). Irregardless, the computational performance of Algorithm 1 on real-world stock indices, as reported in Section 6, is quite attractive.

In the next section, we explore techniques for (a) reliably obtaining high-quality solutions from Algorithm 1 and (b) accelerating Algorithm 1's convergence.

## 5. Improving the performance of the cutting-plane method

### 5.1. Improved upper bounds I: Warm-starts via a discrete ADMM heuristic

The ability of warm-starts to improve the performance of global search strategies such as branch-and-cut is well documented in the literature (see, e.g., Lodi (2013), Bertsimas et al. (2016)). Prompted by this observation, we improve Algorithm 1's performance by injecting high-quality warm-starts. To do so, we assume that $f(\boldsymbol{s})$ is locally Lipschitz continuous with constant $L$ and iteratively solve the following problem, which ranks the differences between each security $i$'s contribution to the portfolio, $x_i$, and its gradient $\nabla_i f(\boldsymbol{s})$:

$$\boldsymbol{s}_{\text{new}} := \arg\min_{\boldsymbol{s}\in S_n^k} ||\boldsymbol{s} - \boldsymbol{x}_{\text{old}} + \frac{1}{L}\nabla f(\boldsymbol{s}_{\text{old}})||_2^2, \tag{21}$$

where we evaluate $\nabla f(\boldsymbol{s}_{\text{old}})$ by solving Problem (3) and invoking Corollary 2, and we find $\boldsymbol{s}_{\text{new}}$ by sorting the residuals by absolute magnitude and selecting the $k$ largest values; this is optimal by Proposition 3 of Bertsimas et al. (2016). We set $L = 10$ as, empirically speaking, this value provides high-quality solutions. We also follow Bertsimas and Van Parys (2016) and Bertsimas et al. (2017) in applying the update rule $\boldsymbol{\alpha}_{t+1} = \frac{t}{t+1}\boldsymbol{\alpha}_t + \frac{1}{t+1}\boldsymbol{\alpha}^*$ to compute the dual variables.

We formalize the warm-start procedure in Algorithm 2, which we apply from 5 distinct and random starting positions to increase the probability that it obtains a high-quality solution apriori. We refer the interested reader to Section 3 of Bertsimas et al. (2016) for an in-depth discussion of discrete first-order heuristics in best subset selection, and Nesterov (2013) for a general theory.

**Algorithm 2** A discrete ADMM heuristic for warm-starts (see Bertsekas (1999)).

**procedure** ADMM HEURISTIC
**input:** $\boldsymbol{X} \in \mathbb{R}^{n\times r}$, $\boldsymbol{y} \in \mathbb{R}^r$, $\boldsymbol{d} \in \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{l} \in \mathbb{R}^m$, $\boldsymbol{u} \in \mathbb{R}^m$,
$\gamma \in \mathbb{R}$, $k \in \{2, ..., n-1\}$, randomRestarts, and $L := 10$
**output:** $\boldsymbol{s}_r \in S_n^k$
  **for** $i = 1$:randomRestarts **do**
    $\boldsymbol{s} \leftarrow \text{rand}(S_n^k)$ ▷ Generate random vector in $S_n^k$
    **while** $\boldsymbol{s}_t != \boldsymbol{s}_{t-1}$ **and** $t < r$ **do**
      $\boldsymbol{\alpha}_t^*, \boldsymbol{\pi}_t^*, \lambda_t^*, \boldsymbol{\beta}_{l,t}^*, \boldsymbol{\beta}_{u,t}^* \leftarrow \text{OptDuals}(\boldsymbol{A}_s,\ \boldsymbol{X}_s,\ \boldsymbol{d}_s)$ ▷ Solve Problem (10)
      $(\boldsymbol{\alpha}^*, \boldsymbol{\pi}^*, \lambda^*, \boldsymbol{\beta}_l^*, \boldsymbol{\beta}_u^*) \leftarrow \frac{1}{t}(\boldsymbol{\alpha}_t^*, \boldsymbol{\pi}_t^*, \lambda_t^*, \boldsymbol{\beta}_{l,t}^*, \boldsymbol{\beta}_{u,t}^*) + \frac{t-1}{t}(\boldsymbol{\alpha}^*, \boldsymbol{\pi}^*, \lambda^*, \boldsymbol{\beta}_l^*, \boldsymbol{\beta}_u^*)$
      $\nabla f(\boldsymbol{s}_t) \leftarrow \frac{-\gamma}{2}(\boldsymbol{X}^\top\boldsymbol{\alpha}^* + \boldsymbol{\pi}^* + \lambda^*\boldsymbol{e} + \boldsymbol{A}^\top(\boldsymbol{\beta}_l^* - \boldsymbol{\beta}_u^*) - \boldsymbol{d})^2$
      $\boldsymbol{x}_t \leftarrow \gamma(\boldsymbol{X}_s^\top\boldsymbol{\alpha}^* + \boldsymbol{A}_s^\top(\boldsymbol{\beta}_l^* - \boldsymbol{\beta}_u^*) + \lambda^*\boldsymbol{e}_s + \boldsymbol{\pi}^* - \boldsymbol{d}_s)$
      $\boldsymbol{s}_{t+1} \leftarrow \arg\min_{\boldsymbol{s}\in S_n^k} ||\boldsymbol{s} - \boldsymbol{x}_t + \frac{1}{L}\nabla f(\boldsymbol{s}_t)||_2^2$, $t \leftarrow t+1$

### 5.2. Improved upper bounds II: Embedding local search within Algorithm 1

The previous subsection describes a warm-start procedure which improves Algorithm 1's performance, by providing a high-quality upper bound apriori. Unfortunately, while this local search method quickly identifies high-quality solutions, it does not exploit global problem structure knowledge and consequently does not identify an optimal solution at scale. This subsection improves upon this situation, by embedding Algorithm 2 within Algorithm 1 to leverage global problem structure knowledge and thereby obtain an optimal solution quickly.

The local-search procedure works as follows: whenever we generate a cut which refines the outer-approximation surface, we receive two pieces of information: the objective function and the subgradient. If the objective function is poor (more than 5% worse than the current incumbent), then an optimal solution is unlikely to be near the current iterate, and we resume branching. Otherwise, motivated by the observation that an optimal solution tends to be surrounded by high-quality solutions, we perform an iteration of Algorithm 2 starting from the solution selected by the cutting-plane procedure. This iteration is effectively a local-search procedure, as the $\boldsymbol{x}_t$ term encourages retaining existing solution components, although we do not impose an explicit search neighbourhood. If the best solution identified by the local-search procedure outperforms the cutting-plane method's current incumbent, then we inject it via a heuristic callback (see CPLEX (2018) or Dunning et al. (2017)).

We also leverage the local search procedure within the heuristic callback, by allowing the solver to request an iteration of Algorithm 2 from a given starting solution, provided the solution's objective value is within 5% of the incumbent. We inject Algorithm 2's solution into the solution pool provided its objective value is within 2.5% of the incumbent; this is a good strategy since optimization solvers leverage techniques such as Relaxation Induced Neighbourhood Search (see Danna et al. (2005)) to obtain new incumbents from high-quality but suboptimal solutions.

### 5.3. Improved lower bounds I: A Second Order Cone relaxation

The preceding subsections describe a procedure which obtains high-quality solutions quickly. In this subsection, we propose a Second Order Cone Problem (SOCP, see Lobo et al. (1998)) which provides a stronger lower bound than that initially obtained by Algorithm 1, although weaker than Algorithm 1's terminal bound. The following theorem provides a derivation of the SOCP:

Theorem 6. *Suppose Assumption 1 holds. Then, the following two optimization problems are equivalent:*

$$\min_{\boldsymbol{s}\in \mathrm{Conv}(S_k^n)} \max_{\substack{\boldsymbol{\pi}\in\mathbb{R}^n_+,\ \boldsymbol{\alpha}\in\mathbb{R}^r,\ \boldsymbol{w}\in\mathbb{R}^n,\\ \boldsymbol{\beta}_l,\ \boldsymbol{\beta}_u\in\mathbb{R}^m_+,\ \lambda\in\mathbb{R}}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}-\frac{\gamma}{2}\sum_i s_i w_i^2+\boldsymbol{y}^\top\boldsymbol{\alpha}+\boldsymbol{\beta}_l^\top\boldsymbol{l}-\boldsymbol{\beta}_u^\top\boldsymbol{u}+\lambda \qquad (22)$$
$$\text{s.t.}\quad \boldsymbol{w}=\boldsymbol{X}^\top\boldsymbol{\alpha}+\boldsymbol{\pi}+\lambda\boldsymbol{e}+\boldsymbol{A}^\top(\boldsymbol{\beta}_l-\boldsymbol{\beta}_u)-\boldsymbol{d}.$$

$$\begin{aligned}\max_{\substack{\boldsymbol{\pi}\in\mathbb{R}^n_+,\ \boldsymbol{\alpha}\in\mathbb{R}^r,\ \boldsymbol{v}\in\mathbb{R}^+_n,\ \boldsymbol{w}\in\mathbb{R}^n,\\ \boldsymbol{\beta}_l,\ \boldsymbol{\beta}_u\in\mathbb{R}^m_+,\ \lambda\in\mathbb{R}}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}+\boldsymbol{y}^\top\boldsymbol{\alpha}+\boldsymbol{\beta}_l^\top\boldsymbol{l}-\boldsymbol{\beta}_u^\top\boldsymbol{u}+\lambda-\boldsymbol{e}^\top\boldsymbol{v}-kt\\ \text{s.t.}\quad & \boldsymbol{w}=\boldsymbol{X}^\top\boldsymbol{\alpha}+\boldsymbol{\pi}+\lambda\boldsymbol{e}+\boldsymbol{A}^\top(\boldsymbol{\beta}_l-\boldsymbol{\beta}_u)-\boldsymbol{d},\\ & v_i\geq\frac{\gamma}{2}w_i^2-t,\quad\forall i\in\{1,...,n\}.\end{aligned}\tag{23}$$

*Proof of Theorem 6* The Sion-Kakutani minimax theorem (see Ben-Tal and Nemirovski (2001)) allows us to exchange the minimum and maximum operators and obtain the following optimization problem:

$$\begin{aligned}\max_{\substack{\boldsymbol{\pi}\in\mathbb{R}^n_+,\ \boldsymbol{\alpha}\in\mathbb{R}^r,\ \boldsymbol{w}\in\mathbb{R}^n,\\ \boldsymbol{\beta}_l,\ \boldsymbol{\beta}_u\in\mathbb{R}^m_+,\ \lambda\in\mathbb{R}}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}+\boldsymbol{y}^\top\boldsymbol{\alpha}+\boldsymbol{\beta}_l^\top\boldsymbol{l}-\boldsymbol{\beta}_u^\top\boldsymbol{u}+\lambda-\frac{\gamma}{2}\max_{\boldsymbol{s}\in\mathrm{Conv}(S^n_k)}\sum_i s_i w_i^2\\ \text{s.t.}\quad & \boldsymbol{w}=\boldsymbol{X}^\top\boldsymbol{\alpha}+\boldsymbol{\pi}+\lambda\boldsymbol{e}+\boldsymbol{A}^\top(\boldsymbol{\beta}_l-\boldsymbol{\beta}_u)-\boldsymbol{d}.\end{aligned}\tag{24}$$

As noted by Zakeri et al. (2014), applying strong duality between the primal problem:

$$\max_{\boldsymbol{s}\in\mathrm{Conv}(S^k_n)}\sum_i s_i w_i^2 := \max\ \boldsymbol{v}^\top\boldsymbol{s}\quad\text{s.t. }\boldsymbol{s}\leq\boldsymbol{e},\ \boldsymbol{e}^\top\boldsymbol{s}=k,\ \boldsymbol{s}\geq\boldsymbol{0},$$

and its dual problem:

$$\min\ \boldsymbol{e}^\top\boldsymbol{u}+kt\text{ s.t. }\boldsymbol{u}+t\boldsymbol{e}\geq\boldsymbol{v},\ \boldsymbol{u}\geq\boldsymbol{0},$$

yields the result. $\square$

Observe that Theorem 6 still holds when the constraint $\sum_i s_i=k$ is relaxed to $\sum_i s_i\leq k$, provided we impose the additional requirement that $t\geq 0$ in Problem (23). Moreover, it is not too hard to see that if all securities are i.i.d. then Problem (23)'s lower bound is exact. This observation motivates us to establish conditions under which Problem (23) provides apriori exact support recovery.

### 5.4. Improved lower bounds II: A verifiable condition for support recovery

The next corollary establishes that the tightness of Problem (23)'s lower bound is sufficient for apriori support recovery.

Corollary 3. **A sufficient condition for support recovery (c.f. Pilanci et al. (2015))** Problem (23)'s lower bound is exact if and only if there exists some $\boldsymbol{s}\in S^k_n$ and set of dual multipliers $(\boldsymbol{v}^*,\boldsymbol{w}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}_l^*,\boldsymbol{\beta}_u^*,\boldsymbol{\pi}^*,\lambda^*)$ which solve Problem (23), such that these two quantities collectively satisfy the following conditions:

$$\begin{aligned}\gamma\sum_i s_i w_i^*=1,\\ \boldsymbol{l}\leq\gamma\sum_i\boldsymbol{A}_i w_i^* s_i\leq\boldsymbol{u},\\ s_i w_i\geq 0,\ \forall i,\\ v_i^*=0,\ \forall i\text{ s.t. }s_i=0.\end{aligned}\tag{25}$$

*Proof of Corollary 3* Let there exist some set of dual multipliers $(\boldsymbol{v}^*, \boldsymbol{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}_l^*, \boldsymbol{\beta}_u^*, \boldsymbol{\pi}^*, \lambda^*)$ which solve Problem (23), and binary vector $\boldsymbol{s} \in S_n^k$, such that these two quantities collectively satisfy the conditions encapsulated in Expression (25). Then, this optimal solution to Problem (23) provides the following lower bound for Problem (1):

$$-\frac{1}{2}\boldsymbol{\alpha}^{*\top}\boldsymbol{\alpha}^* + \boldsymbol{y}^\top\boldsymbol{\alpha}^* + \boldsymbol{\beta}_l^{*\top}\boldsymbol{l} - \boldsymbol{\beta}_u^{*\top}\boldsymbol{u} + \lambda^* - \boldsymbol{e}^\top\boldsymbol{v}^* - kt^*.$$

Moreover, let $\hat{\boldsymbol{x}}$ be a candidate solution to Problem (1) defined by $\hat{x}_i := \gamma w_i s_i$. Then, $\hat{\boldsymbol{x}}$ is feasible for Problem (1), since $\boldsymbol{l} \leq A\hat{\boldsymbol{x}} \leq \boldsymbol{u}$, $\boldsymbol{e}^\top\hat{\boldsymbol{x}} = 1$, $\hat{\boldsymbol{x}} \geq \boldsymbol{0}$ and $||\hat{\boldsymbol{x}}||_0 \leq k$ by Expression (25) and the definition of $\boldsymbol{s}$. Moreover, the construction of $\hat{\boldsymbol{x}}$ via the dual problem's KKT conditions imply that Problem (1)'s objective when $\boldsymbol{x} = \hat{\boldsymbol{x}}$ is given by:

$$-\frac{1}{2}\boldsymbol{\alpha}^{*\top}\boldsymbol{\alpha}^* + \boldsymbol{y}^\top\boldsymbol{\alpha}^* + \boldsymbol{\beta}_l^{*\top}\boldsymbol{l} - \boldsymbol{\beta}_u^{*\top}\boldsymbol{u} + \lambda^* - \frac{1}{2\gamma}\hat{\boldsymbol{x}}^\top\hat{\boldsymbol{x}},$$

which is less than or equal to Problem (23)'s objective, since $v_i^* = 0 \ \forall i$ s.t. $s_i = 0$.

For the converse, if no optimal set of dual multipliers satisfies Expression (25), then there is no feasible $\boldsymbol{x} := \gamma\boldsymbol{w}$ which provides the same objective value as Problem (23)'s lower bound, and therefore the SOCP gap is non-zero. □

A natural way of selecting an $\boldsymbol{s}$ which is likely to satisfy Expression (25) is to set $s_i = 1$ when $\gamma(w_i^*)^2 > 2t$, since $s_i$ is the dual multiplier for the constraint $v_i \geq \frac{\gamma}{2}w_i^2 - t$ in Problem (23). However, we do not pursue this issue, as our cutting-plane method's computational properties are already quite attractive, and Expression (25) is not universally satisfiable.

### 5.5. An improved cutting-plane method

Combining the analysis in this section with Algorithm 1 leads to Algorithm 3, which, by Theorem 5, also converges to an optimal sparse portfolio $\boldsymbol{x}^*$ in a finite number of iterations. Moreover, by applying the embedded local-search procedure it typically identifies the optimal solution more rapidly than Algorithm 1, and by applying the SOCP lower bound, it typically prunes a higher proportion of nodes in the branch-and-cut tree than Algorithm 1.

For simplicity, we do not depict the heuristic callback described in the previous subsection within this procedure, although we include it in our code. We also apply the SOCP lower bound by adding a lazy constraint of the form $\theta \geq \text{ofv}_{\text{SOCP}}$ after generating 100 cuts, where '$\text{ofv}_{\text{SOCP}}$' refers to Problem (23)'s optimal objective. We do not introduce the lower bound immediately, because it is uninformative and the solver makes lower quality branching decisions when it is binding in the master problem. This approach allows the local-search heuristic described in the previous subsection to identify a (near) optimal solution before the uninformative lower bound is applied.

**Algorithm 3** A refined cutting-plane method for Problem (1).

**procedure** REFINED CUTTING-PLANE METHOD
**input:** $\boldsymbol{X} \in \mathbb{R}^{r\times n}$, $\boldsymbol{y} \in \mathbb{R}^r$, $\boldsymbol{d} \in \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{l} \in \mathbb{R}^m$, $\boldsymbol{u} \in \mathbb{R}^m$, $\gamma \in \mathbb{R}$ and $k \in \{2,...,n-1\}$
**output:** $\boldsymbol{s}^* \in S_n^k$ and $\boldsymbol{x}^* \in \mathbb{R}^n$
  $\boldsymbol{s}_1 \leftarrow$ warm start, $t \leftarrow 1$ ▷ See Algorithm 2
  Add cut $\theta \geq f(\boldsymbol{s}_1) + \nabla f(\boldsymbol{s}_1)^\top(\boldsymbol{s} - \boldsymbol{s}_1)$
  $\text{ofv}_{\text{SOCP}} \leftarrow$ MaxValue(Problem (23)) ▷ See Theorem 6
  **while** $\theta_t < f(\boldsymbol{s}_t)$ **do**
    $\boldsymbol{\alpha}_t^*, \boldsymbol{\pi}_t^*, \lambda_t^*, \boldsymbol{\beta}_{l,t}^*, \boldsymbol{\beta}_{u,t}^* \leftarrow \text{OptDuals}(\boldsymbol{A}_s,\ \boldsymbol{X}_s,\ \boldsymbol{d}_s)$
    **if** Dual Unbounded **then** ▷ Dual always feasible.
      Add feasibility cut $\sum_i \hat{s}_i(1-s_i) + \sum_i (1-\hat{s}_i)s_i \geq 1$
    **else**
      $f(\boldsymbol{s}_t) \leftarrow -\frac{\gamma}{2}||\boldsymbol{X}_s^\top \boldsymbol{\alpha}_t^* + \lambda_t^* \boldsymbol{e}_s + \boldsymbol{\pi}_t^* + \boldsymbol{A}_s^\top(\boldsymbol{\beta}_{l,t}^* - \boldsymbol{\beta}_{u,t}^*) - \boldsymbol{d}_s||_2^2$
      $\qquad -\frac{1}{2}||\boldsymbol{\alpha}_t^*||_2^2 + \boldsymbol{y}^\top \boldsymbol{\alpha}_t^* + \boldsymbol{l}^\top \boldsymbol{\beta}_{l,t}^* - \boldsymbol{u}^\top \boldsymbol{\beta}_{u,t}^* + \lambda_t^*$
      **for** $i \in \{1,...,n\}$ **do**
        $\nabla f(\boldsymbol{s}_t)_i \leftarrow -\frac{\gamma}{2}(\boldsymbol{X}_{\boldsymbol{i}}^\top \boldsymbol{\alpha}_t^* + \boldsymbol{A}_i^\top(\boldsymbol{\beta}_{l,t}^* - \boldsymbol{\beta}_{u,t}^*) + \lambda_t^* + \pi_{i,t}^* - d_i)^2$
    **if** $f(\boldsymbol{s}_t) \leq 1.05 f(\boldsymbol{s}_{\text{best}})$ **then**
      $\boldsymbol{s}_{l,t} \leftarrow \text{localSearch}(\boldsymbol{s}_t)$ (See Algorithm 2)
      **if** $f(\boldsymbol{s}_{l,t}) < f(\boldsymbol{s}_{\text{best}})$ **then**
        Inject $\boldsymbol{s}_{l,t}$ into solution pool
    **if** $t = 100$ **then**
      Add optimality cut $\theta \geq \text{ofv}_{\text{SOCP}}$
    $\boldsymbol{s}_{t+1}, \theta_{t+1} \leftarrow \arg\min_{\boldsymbol{s},\theta} \left\{\theta \text{ s.t. } \boldsymbol{s} \in S_n^k,\ \theta \geq f(\boldsymbol{s}_j) + \nabla f(\boldsymbol{s}_j)^\top(\boldsymbol{s} - \boldsymbol{s}_j),\ \forall j \in \{1,...,t\},\ \text{Feasibility cuts}\right\}$
    $t \leftarrow t+1$
  $\boldsymbol{s}^* \leftarrow \boldsymbol{s}_t,\ \boldsymbol{x}^* \leftarrow \boldsymbol{0},\ \boldsymbol{x}_s^* \leftarrow \gamma(\boldsymbol{X}_s^\top \boldsymbol{\alpha}_t^* + \boldsymbol{A}_s^\top(\boldsymbol{\beta}_{l,t}^* - \boldsymbol{\beta}_{u,t}^*) + \lambda_t^* \boldsymbol{e}_s + \boldsymbol{\pi}_t^* - \boldsymbol{d}_s)$

In the next section, we explore Algorithm 3's performance on real-world data.

## 6. Computational experiments on real-world data

In this section, we present a sequence of numerical experiments using the Julia programming language (see Bezanson et al. (2017)). All optimization problems are formulated using the JuMP package version 0.18.0 (see Dunning et al. (2017)), and, unless stated otherwise, solved using CPLEX (2018) version 12.8.0. Moreover, all computations are performed on the *engaging* cluster, a high performance cluster at the *Massachusetts Green High Performance Computing Centre* (MGH-PCC) which comprises a set of Intel Xeon E5 − 2600 v4 2.0GHz processors. All computations are allocated 1 virtual CPU core and 32 GB RAM. The big-M formulations in Section 6.1 are solved using 28 threads. However, as JuMP is currently not thread-safe and CPLEX currently cannot combine multiple threads with lazy callbacks and non-thread-safe code, we provide Algorithm 3 with a single thread (this situation could be remedied by implementing Algorithm 3 in C++, but our Julia implementation already performs quite attractively).

In the following numerical experiments, we solve the following optimization problem, which places the multiplier $\alpha$ on the return term but is mathematically equivalent to Problem (1):

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}+\frac{1}{2\gamma}||\boldsymbol{x}||_2^2-\alpha\boldsymbol{\mu}^\top\boldsymbol{x} \text{ s.t. } \boldsymbol{l}\leq\boldsymbol{A}\boldsymbol{x}\leq\boldsymbol{u},\ \boldsymbol{e}^\top\boldsymbol{x}=1,\ \boldsymbol{x}\geq\boldsymbol{0},\ ||\boldsymbol{x}||_0\leq k. \tag{26}$$

We aim to answer the following three questions:

1. How does Algorithm 3 compare to existing state-of-the-art solution methods?
2. How does Algorithm 3 scale as a function of the number of securities in the buyable universe?
3. How sensitive are optimal solutions to Problem (1) to the hyperparameters $\alpha, \gamma, k$?

### 6.1. A comparison between Algorithm 3 and existing state-of-the-art methods

We now present a direct comparison of Algorithm 3 with CPLEX (2018) version 12.8.0 and Gurobi (2018) version 7.5.1, where both solvers are provided with the big-M formulation outlined in Problem (3). As both solvers implement most techniques outlined in our summary of existing approaches, including the lifted branch-and-cut approach proposed by Vielma et al. (2017), this experiment can reasonably be viewed as a comparison between our method and the state-of-the-art.

Table 2 depicts the time required for all 3 approaches to determine an optimal allocation of funds between $k$ securities for a given choice of robustness parameters, with the problem data corresponding to the 5 mean-variance portfolio optimization problems described by Chang et al. (2000) and subsequently included in the OR-library test set collated by Beasley (1990). The experimental results illustrate that our approach is several orders of magnitude more efficient on all problem instances considered, and that this efficiency improvement allows us to solve portfolio optimization problems which are currently considered to be intractable.

A consequence of our method's performance is that we can construct provably optimal cardinality-constrained efficient frontiers for the OR library instances. To this end, Figure 1 depicts the frontiers for the OR library problems *port1*—*port5*, with the robustness parameter $\gamma=\frac{1}{\sqrt{n}}$, and the portfolio variances computed using the regularized covariance matrix $\frac{1}{\gamma}\mathbb{I}+\Sigma$.

Our main findings from this set of experiments are as follows:

1. Existing approaches to Problem (1) do not scale to real-world problem sizes because big-M formulations provide low-quality lower bounds which cause solvers to be overly optimistic when expanding nodes and lead to branch-and-bound trees with $3-5$ orders of magnitude more nodes expanded than necessary. Indeed, for the OR library problems where CPLEX and Gurobi could not verify optimality, both solvers provided significantly weaker lower bounds than the SOCP lower bound, even after expanding $10^7$ nodes, although their incumbent solutions are near-optimal.

2. The solve times for high-return portfolios are smaller than minimum-variance portfolios, and the later problems are significantly more ill-conditioned, particularly for problem *port5*.

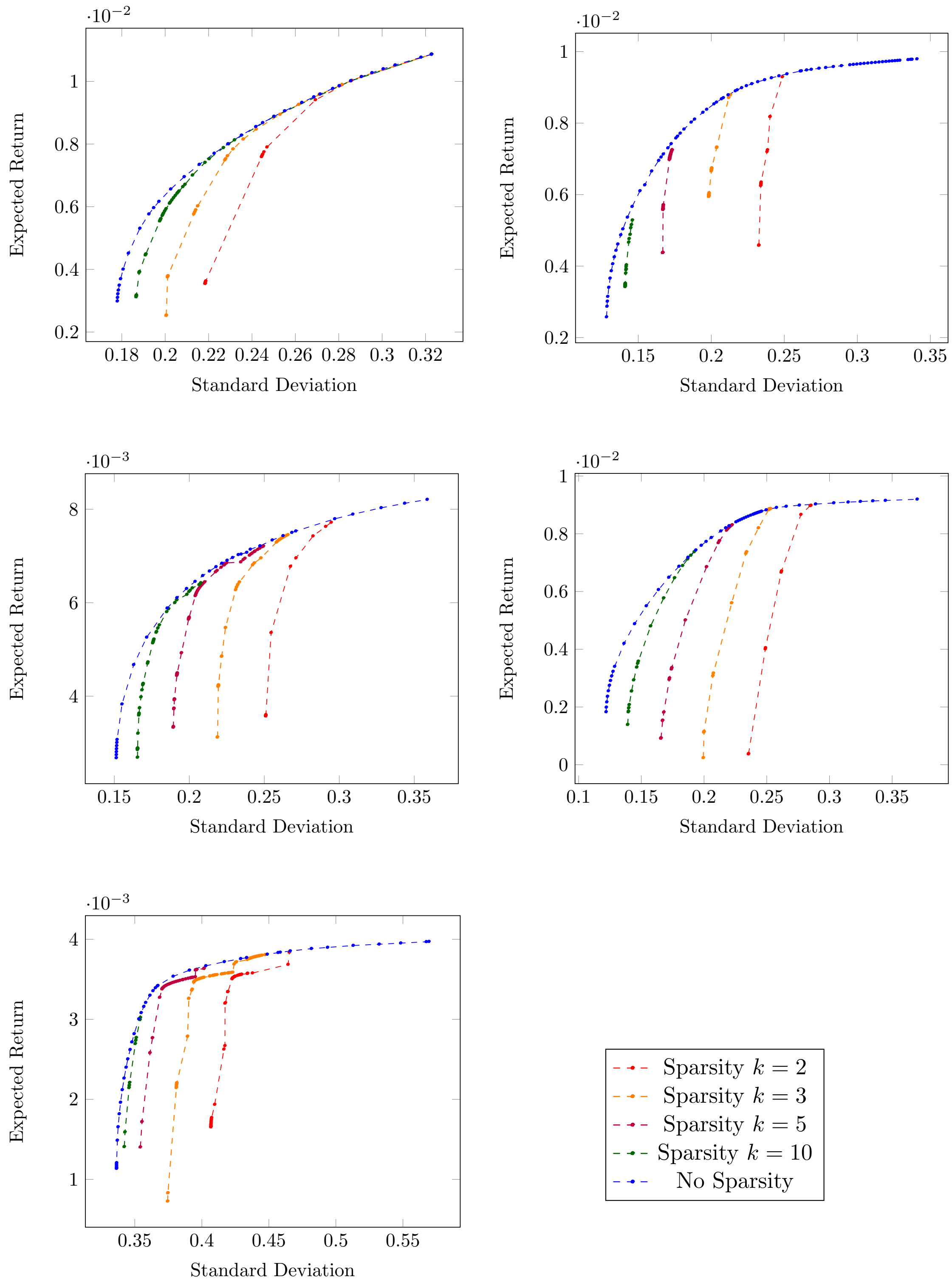
**Table 2** Mean runtime in seconds per approach. We impose a maximum runtime of $36,000$s an optimality tolerance of $10^{-12}$ for Algorithm 3 and CPLEX, $10^{-8}$ for Gurobi, and robustness parameters $\gamma = \frac{1}{\sqrt{n}}$, $\alpha = 0.5$.

| Problem | Rank($\boldsymbol{\Sigma}$) | No. securities | Subset | Runtime (s) | | |
|---|---|---|---|---|---|---|
| | | Available | Size | Algorithm 3 | CPLEX Big-M | Gurobi Big-M |
| port 1 | 31 | 31 | 5 | 4.66 | 858.00 | 51.54 |
| | | | 10 | 0.59 | 7,752.97 | 10,071.33 |
| | | | 20 | 0.67 | 9,805.82 | 18,424.81 |
| port 2 | 85 | 85 | 5 | 6.51 | 18,828.37 | 35,878.19 |
| | | | 10 | 1.67 | 36,000.00 | 36,000.00 |
| | | | 20 | 1.75 | 36,000.00 | 36,000.00 |
| port 3 | 89 | 89 | 5 | 5.67 | 28,423.95 | 36,000.00 |
| | | | 10 | 1.64 | 36,000.00 | 36,000.00 |
| | | | 20 | 1.83 | 36,000.00 | 36,000.00 |
| port 4 | 98 | 98 | 5 | 23.26 | 36,000.00 | 36,000.00 |
| | | | 10 | 10.46 | 36,000.00 | 36,000.00 |
| | | | 20 | 12.02 | 36,000.00 | 36,000.00 |
| port 5 | 225 | 225 | 5 | 91.27 | 36,000.00 | 36,000.00 |
| | | | 10 | 69.76 | 36,000.00 | 36,000.00 |
| | | | 20 | 72.24 | 36,000.00 | 36,000.00 |

3. For high-return portfolios, Problem (23) provides an exact lower bound, while the SOCP gap tends to be small but non-zero for low-variance low-return sparse portfolios. Consequently, the problem difficulty depends on both $\gamma$ and $\alpha$, with exact solutions returned quickly when $\gamma \leq \frac{100}{\sqrt{n}}$ and $\alpha \geq 0.1$, and provably high-quality (e.g., a bound gap $< 2\%$) solutions returned quickly otherwise.
4. When selecting sparse portfolios from a buyable universe of 225 or fewer securities, including both the mean and variance objective components in the objective does not identify most non-dominated portfolios, due to integrality, and more sophisticated multi-objective techniques such as the $\epsilon$-constraint method (see Ehrgott (2005)), which optimizes $\boldsymbol{\mu}^\top \boldsymbol{x}$ subject to an upper bound on $\boldsymbol{x}^\top \left(\frac{1}{\gamma}\mathbb{I} + \boldsymbol{\Sigma}\right)\boldsymbol{x}$, are necessary to generate sparse non-dominated frontiers. However, the next sequence of experiments demonstrates that this is not the case when selecting from larger universes of assets, and therefore we do not modify Algorithm 3.

For the rest of the paper, we do not consider solving big-M formulations with either CPLEX or Gurobi, as these formulations do not scale to larger problem sizes with 500 or more securities in the universe of buyable assets.

**Figure 1** Efficient frontiers by cardinality for the OR library problems *port1*—*port5* (see Chang et al. (2000)). The problem data corresponds to: 31 securities in the Hang Seng index (*port1*, top left), 85 securities in the DAX 100 index (*port2*, top right), 89 securities in the FTSE 100 (*port3*, middle left), 98 securities in the S& P 100 (*port4*, middle right), and 225 assets in the Nikkei 225 (*port5*, bottom left).



### 6.2. Exploring the scalability of Algorithm 3

In this section, we explore Algorithm 3's scalability with respect to the number of securities in the buyable universe, by solving several large-scale sparse portfolio selection problems to provable optimality: the S&P 500, the Russell 1000, and the Wilshire 5000.

Our first experiment measures the time required to construct cardinality-constrained portfolios for these three stock indices. In all three cases, the problem data is taken from daily closing prices from January 3 2007 to December 29 2017, which are obtained from Yahoo! Finance via the R package *quantmod* (see Ryan et al. (2018)), and rescaled to correspond to a holding period of one month. We apply Singular Value Decomposition to obtain low-rank estimates of the correlation matrix, and rescale the low-rank correlation matrix by each asset's variance to obtain a low-rank covariance matrix $\boldsymbol{\Sigma}$. We also omit days with a greater than 20% change in closing prices when computing the mean and covariance for the Russell 1000 and Wilshire 5000, since these changes occur on low-volume trading and typically reverse the next day. Tables 3—5 depict the time required to solve the problem to provable optimality for different choices of $\gamma$, $k$, and rank($\Sigma$). For $\gamma = \frac{100}{\sqrt{n}}$ the problem becomes more difficult and Algorithm 3 does not close the bound gap to 0 for the S&P 500 or the Russell 1000, and we instead provide the magnitude of the bound gap at $36,000$s (note that Algorithm 3 typically obtains its lower bound from the SOCP when $\gamma = \frac{100}{\sqrt{n}}$, meaning we could obtain a similar bound gap within $3,600$s or less). For the Wilshire 5000, CPLEX could not solve the QOs or SOCPs in a reasonable amount of time, and therefore we instead used the solver Mosek (2010) to solve these problems; doing so reduced the time required to verify optimality compared to the Russell 1000, although we found that CPLEX was more numerically stable.

**Table 3** Algorithm 3's average solution time in seconds for the S&P500, with a maximum runtime of $36,000$s, an optimality tolerance of $10^{-12}$, and $\alpha = 1/15$. For $\gamma = \frac{100}{\sqrt{n}}$, we provide the bound gap at $36,000$s.

| Robustness | Subset | Runtime (s) by Rank($\boldsymbol{\Sigma}$) | | | | |
|---|---|---|---|---|---|---|
| | Size | 50 | 75 | 100 | 150 | 200 |
| $\gamma = 1/\sqrt{n}$ | 50 | 274.00 | 164.00 | 146.26 | 245.17 | 374.18 |
| | 100 | 134.35 | 197.69 | 155.88 | 190.87 | 253.19 |
| | 150 | 208.57 | 175.37 | 192.43 | 257.66 | 221.58 |
| | 200 | 233.17 | 207.01 | 169.50 | 298.39 | 297.28 |
| $\gamma = 10/\sqrt{n}$ | 50 | 105.20 | 152.76 | 144.28 | 181.90 | 344.49 |
| | 100 | 193.96 | 146.31 | 177.75 | 167.95 | 277.72 |
| | 150 | 229.71 | 172.80 | 175.58 | 244.53 | 297.11 |
| | 200 | 262.99 | 157.42 | 172.80 | 205.25 | 370.54 |
| $\gamma = 100/\sqrt{n}$ | 50 | 0.072 % | 0.041 % | 0.039 % | 0.024 % | 0.016 % |
| | 100 | 0.048 % | 0.033 % | 0.036 % | 0.011 % | 0.007 % |
| | 150 | 0.032 % | 0.002 % | 1.157 % | 0.004 % | 0.011 % |
| | 200 | 0.016 % | 0.010 % | 0.012 % | 0.006 % | 0.004 % |

**Table 4** Algorithm 3's average solution time in seconds for the Russell 1000. We impose the same experimental parameters as Table 3.

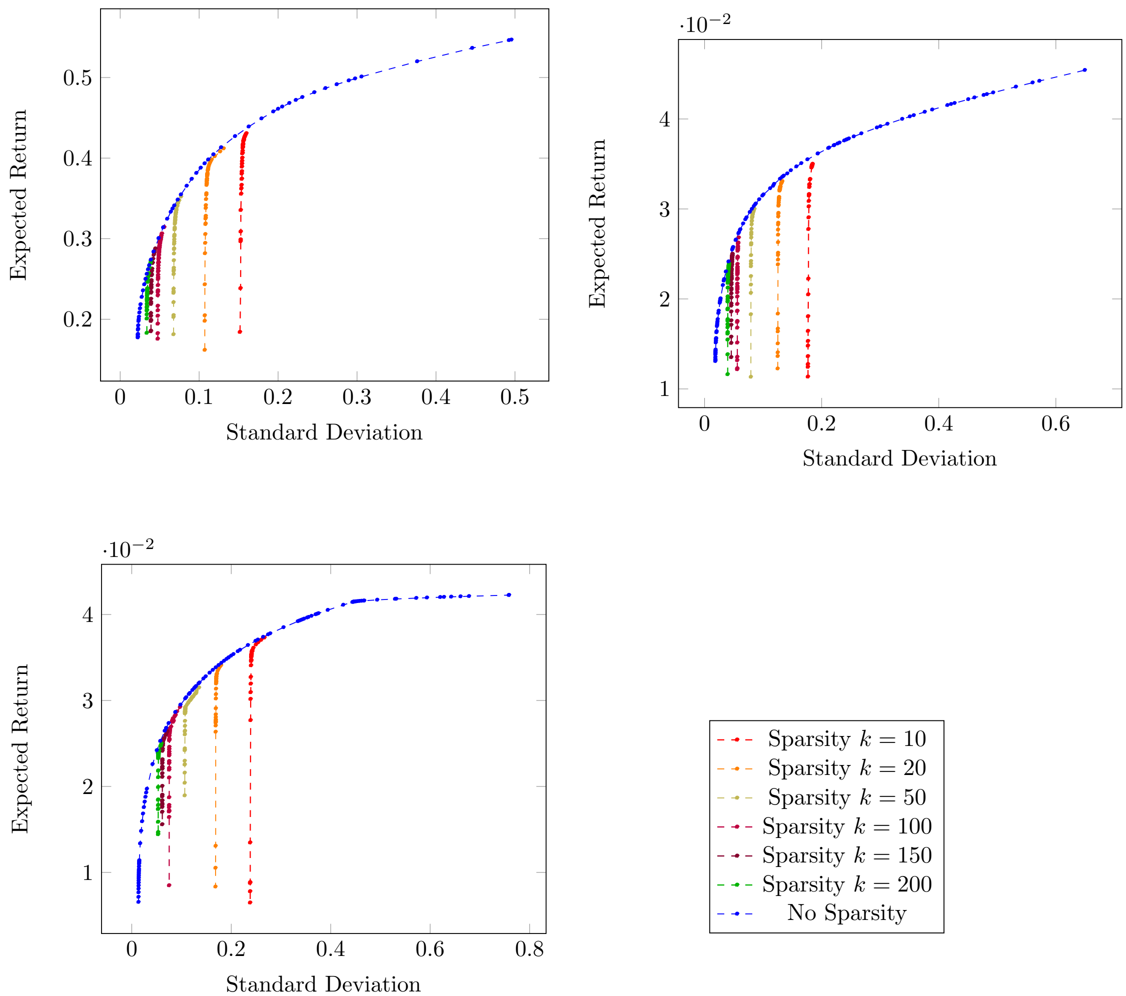
| Robustness | Subset | Runtime (s) by Rank($\boldsymbol{\Sigma}$) | | | | | |
|---|---|---|---|---|---|---|---|
| | Size | 50 | 100 | 150 | 200 | 250 | 300 |
| $\gamma = 1/\sqrt{n}$ | 50 | 1,732.27 | 2,528.32 | 1,981.23 | 2,321.14 | 2785.64 | 3,167.70 |
| | 100 | 2,240.39 | 3,257.74 | 3,815.93 | 3,815.11 | 3,935.69 | 3,712.93 |
| | 150 | 2,646.22 | 3,717.27 | 3,984.35 | 3,816.73 | 4,156.78 | 4,063.61 |
| | 200 | 3,647.23 | 3,924.73 | 4,068.87 | 4,130.14 | 4,237.62 | 4,459.85 |
| $\gamma = 10/\sqrt{n}$ | 50 | 2,676.58 | 2,259.79 | 1,929.42 | 3,824.94 | 2,711.82 | 3,320.07 |
| | 100 | 1,806.98 | 3,462.54 | 4,038.32 | 3,943.21 | 3,893.24 | 4,767.94 |
| | 150 | 3,057.83 | 3,737.85 | 3,757.46 | 3,837.42 | 3,970.30 | 4,181.72 |
| | 200 | 3,763.18 | 3,805.41 | 4,009.27 | 4,165.68 | 4,245.00 | 4,642.60 |
| $\gamma = 100/\sqrt{n}$ | 50 | 0.116% | 0.154% | 0.267% | 0.160% | 0.288% | 0.113% |
| | 100 | 0.421% | 0.002% | 0.137% | 0.010% | 0.041% | 0.036% |
| | 150 | 0.014% | 0.403% | 0.184% | 0.093% | 0.025% | 0.057% |
| | 200 | 0.003% | 0.010% | 0.025% | 0.075% | 0.029% | 0.009% |

**Table 5** Algorithm 3's average solution time in seconds for the Wilshire 5000. We impose the same experimental parameters as Table 3.

| Robustness | Subset | Runtime (s) by Rank($\boldsymbol{\Sigma}$) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Size | 50 | 100 | 200 | 300 | 400 | 500 | 1000 |
| $\gamma = 1/\sqrt{n}$ | 50 | 8.46 | 15.4 | 319.44 | 517.03 | 1,836.66 | 2,350.97 | 36,000.00 |
| | 100 | 32.32 | 22.28 | 27.68 | 60.29 | 65.29 | 93.88 | 2511.79 |
| | 150 | 11.54 | 21.67 | 35.62 | 43.03 | 43.75 | 48.87 | 88.14 |
| | 200 | 17.95 | 39.25 | 40.74 | 43.24 | 57.87 | 65.27 | 101.53 |
| $\gamma = 10/\sqrt{n}$ | 50 | 40.77 | 40.86 | 330.38 | 406.34 | 2093.62 | 2457.64 | 36,000.00 |
| | 100 | 18.35 | 15.79 | 24.91 | 39.75 | 57.27 | 82.25 | 2130.31 |
| | 150 | 13.4 | 21.24 | 37.37 | 43.56 | 45.36 | 49.92 | 80.79 |
| | 200 | 19.1 | 33.83 | 40.06 | 46.55 | 57.41 | 104.11 | 100.23 |
| $\gamma = 100/\sqrt{n}$ | 50 | 32.99 | 38.26 | 313.87 | 2,027.49 | 2,803.79 | 6,213.71 | 36,000.00 |
| | 100 | 17.75 | 40.05 | 34.75 | 46.33 | 61.15 | 95.63 | 14,499.67 |
| | 150 | 39.08 | 257.65 | 91.04 | 142.45 | 718.61 | 156.41 | 207.6 |
| | 200 | 23,451.17 | 30,049.62 | 20,591.48 | 30,066.12 | 29,046.59 | 30,079.99 | 36,000.00 |

Our second experiment constructs cardinality-constrained efficient frontiers for the same three stock indices, with the same problem data. We used the rank-200 approximation of the covariance matrix for the S&P 500, the rank-300 approximation of the covariance matrix for the Russell 1000, and the rank-1000 approximation of the covariance matrix for the Wilshire 5000. Figure 2 depicts the corresponding cardinality-constrained efficient frontiers for $\gamma = \frac{100}{\sqrt{n}}$, with $k$ and $\alpha$ varying.

**Figure 2** Efficient frontiers by cardinality for: the S&P 500 (top left), the Russell 1000 (top right), and the Wilshire 5000 (bottom left). The S&P 500 corresponds to a holding period of one year, and the other two plots correspond to a holding period of one month.




Our main findings from this set of experiments are as follows:

1. The optimality gap depends heavily on $\alpha$ for all three problems, with problems where $\alpha > 0.05$ consistently solved within the timelimit, problems where $0.05 > \alpha > 0.01$ solved to a bound gap of 0.05%, and problems where $\alpha < 0.01$ solved to a bound gap of $0.5 - 2\%$. In all three cases, combining the SOCP lower bound with the best solution found after running Algorithm 3 for 500 seconds yields a solution provably within $1 - 2\%$ of optimality.

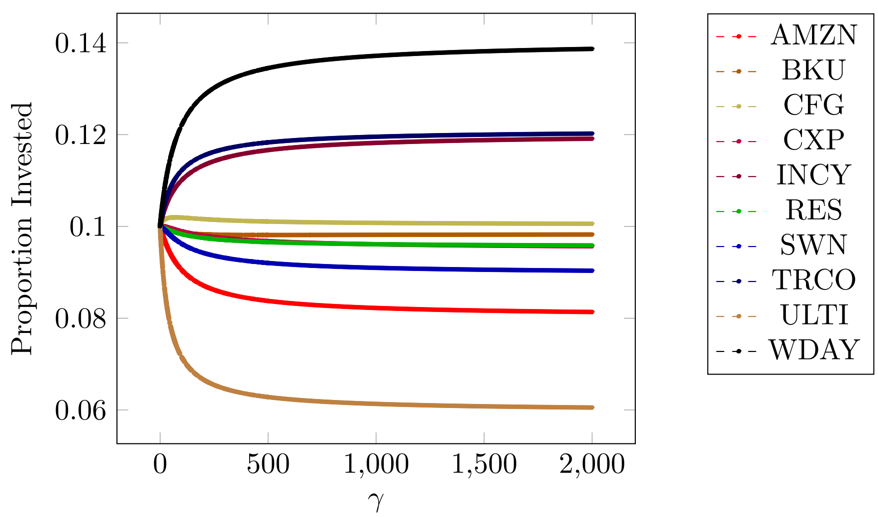
2. Unlike the previous set of experiments, our approach of including both the mean and variance terms in the objective with different values of the multiplier $\alpha$ identifies most non-dominated (with respect to mean-variance) portfolios on the cardinality-constrained efficient frontier, and multi-objective techniques are not needed.

### 6.3. Exploring Problem (1)'s sensitivity to its hyperparameters

Our next sequence of experiments explores Problem (1)'s stability to changes in its hyperparameters $\gamma$, $k$ and $\alpha$. The first experiment studies $\boldsymbol{x}^*$'s sensitivity to $\gamma$ for a rank$-300$ approximation of the Russell 1000 with a one month holding period, a cardinality budget $k = 10$ and a robustness budget of $\alpha = 1/5$.

Figure 3 depicts the relationship between $\boldsymbol{x}^*$ and $\gamma$ for this set of hyperparameters, and indicates that $\boldsymbol{x}^*$ is stable with respect to small changes in $\gamma$. Interestingly, the best choice of $\boldsymbol{s}$ found with a time limit of $36,000$ seconds does not change, for all values of $\gamma$. Moreover, Algorithm 3 certifies optimality almost immediately for small values of $\gamma$, and injecting an optimal solution for a small value of $\gamma$ as a warm-start yields an apriori bound gap of 0.1% for larger values of $\gamma$. This phenomenon suggests that when $\gamma$ is large and $k$ is small, a good strategy comprises solving Problem (1) for $\gamma = \frac{1}{\sqrt{n}}$ and using this solution as a warm-start for Algorithm 3.

**Figure 3** Sensitivity to $\gamma$ for the Russell 1000 with $\alpha = 1/5$. The optimal stock indices $\boldsymbol{s}^*$ did not change for any $\gamma$, and the SOCP bound gap was less than 0.1% for all $\gamma$.

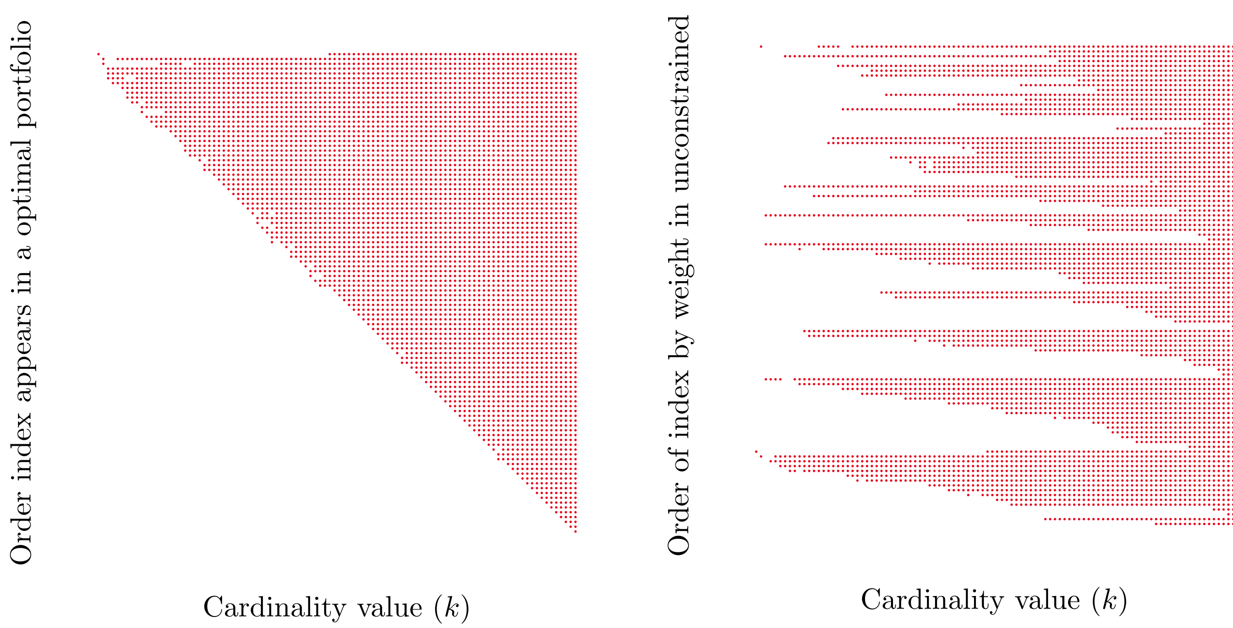



Our second experiment studies Problem (1)'s sensitivity to changes in the cardinality budget $k$ with a robustness budget $\gamma = \frac{1}{\sqrt{n}}$, $\alpha = 1/5$ and the same problem data as the previous experiment. In this experiment, incrementing the cardinality constraint results in the optimal allocation of funds $\boldsymbol{x}^*$ changing whenever the cardinality constraint is binding. Therefore, we consider changes in $\boldsymbol{s}^*$ rather than $\boldsymbol{x}^*$ when performing the sensitivity analysis, and take the view that Problem (1) is stable with respect to changes in $k$ if $\boldsymbol{s}^*$ does not change too much. This is a reasonable perspective when changes in $k$ correspond to investing funds from a new investor.

To this end, we compute the optimal stock indices $i: s_i^* = 1$ for each $k \in \{1, ..., 100\}$ and plot the sparsity patterns against (a) the order in which stock indices are first selected in an optimal solution as we increase $k$, and (b) the weight which the 100 securities selected in (a) are allocated in a continuous relaxation of Problem (1) which does not have a cardinality constraint. In the resulting plots, a strictly upper diagonal matrix would indicate that incrementing $k$ by 1 results in the same securities selected as for an optimal $k$-sparse portfolio, plus one new security. Figure 4 depicts the resulting sparsity pattern, and suggests that the heuristic of ranking securities by the order in which they first appear in a sparsity pattern is near-optimal (since the left matrix is very nearly strictly upper triangular), while ranking securities according to their weight in the unconstrained problem is far from optimal (since the right matrix is clearly not upper triangular).

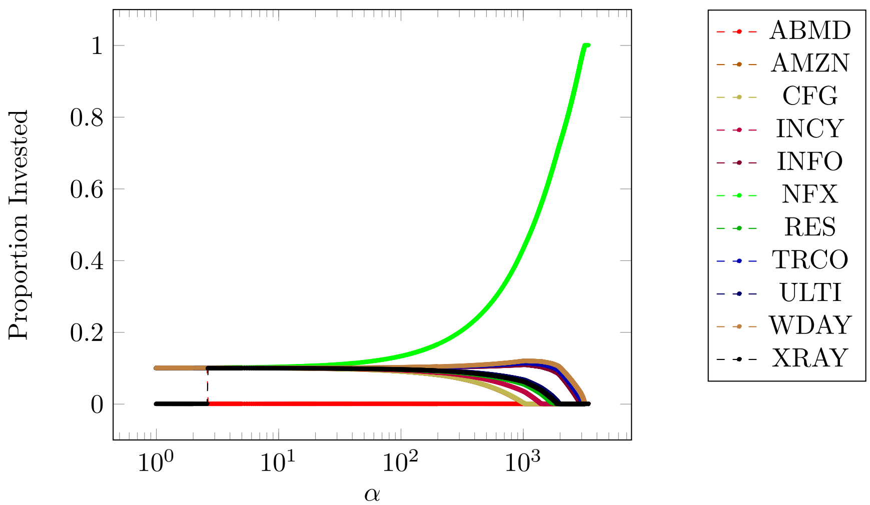
This observation provides a connection to the work of Koç and Morton (2014), who study the problem of generating ranked-list policies for similar problems with a stochastic cardinality budget $k$. Indeed, the stability of the optimal indices indicates that the policy of ranking securities by the order they are first seen in a $k$-sparse portfolio for increasing $k$ is essentially optimal for the ranked-list problem. Consequently, investors can generate approximate ranked-list policies by solving Problem (1) for each $k$, and comparing their list with the securities recently purchased by competitors: if unusually large quantities of securities far down the list were recently purchased, then the competitor may posses private information.

**Figure 4** Sparsity pattern by $k$ for the Russell 1000 with $\alpha = 1/5$, $\gamma = \frac{1}{\sqrt{n}}$, sorted by: the order the indices first appear in an optimal solution (left), the weight assigned without a sparsity constraint (right).



Our third experiment studies $\boldsymbol{x}^*$'s sensitivity to changes in the uncertainty budget $\alpha$ with a fixed robustness budget of $\gamma = \frac{1}{\sqrt{n}}$, cardinality budget of $k = 10$, and the same problem data as the previous experiment. Figure 5 depicts the resulting allocation of funds for $\alpha \geq 1$; we do not depict the allocation of funds for $\alpha < 1$ as this comprises allocating $0.1 \pm 0.001$ to each security selected for various subsets $\boldsymbol{s}$.

**Figure 5** Sensitivity to $\alpha$ for a rank$-300$ approximation of the Russell 1000 with $\gamma = 1/\sqrt{n}$.




### 6.4. Summary of findings from numerical experiments

After performing the numerical experiments discussed in the preceding subsections, we are in a position to answer the three questions introduced at the start of this section.

Our findings are as follows:

1. Algorithm 3 is $3-4$ orders of magnitude more efficient than state-of-the-art MIQO solvers such as Gurobi and CPLEX for the smaller real-world problems in the OR-library test-set. This efficiency improvement is due to both (a) our ability to generate stronger and more informative lower bounds via dual subproblems and SOCP bounds, and (b) our ability to exploit our subgradient representation to obtain and polish high-quality solutions.

2. Algorithm 3 scales to solve real-world problem instances which comprise selecting assets from universes with $1,000$s of securities, such as the Russell 1000 and the Wilshire 5000, while existing state-of-the-art approaches such as CPLEX and Gurobi do not scale to these problem sizes, because they cannot attain sufficiently strong lower bounds and therefore their branch-and-bound trees quickly become too large to fit into memory.

3. Solutions to Problem (1) are stable with respect to the hyperparameters $\alpha$ and $\gamma$. Moreover, while for small values of $k$ optimal solutions are unstable to changes in the cardinality budget, for

$k \geq 20$ the optimal indices for a $(k+1)$-sparse portfolio typically correspond to those for a $k$-sparse portfolio, plus one additional security.

## 7. Conclusions

This paper describes a scalable algorithm for solving quadratic optimization problems subject to sparsity constraints, and applies it to the problem of sparse portfolio selection. Although sparse portfolio selection is NP-hard, and therefore considered to be intractable, our algorithm provides provably optimal portfolios even with the number of securities in the $1,000$s.

Our work is related to that of Balas et al. (1993) and Frangioni and Gentile (2006), who both recognized that big-M based MINLP formulations provide weak relaxations, and respectively proposed lift-and-project and perspective cuts to obtain tighter relaxations. Notably however, the formulations presented here provide similarly tight relaxations without an explosion in the number of decision variables, which makes them quite computationally attractive.

A straightforward extension to our approach comprises replacing the linear inequality constraints with more general constraints, such as the chance and buy-in threshold constraints studied by Bonami and Lejeune (2009). This extension requires a more general notion of duality, such as conic duality (see Ben-Tal and Nemirovski (2001)), but otherwise follows directly from our approach.

#### Acknowledgements

We are grateful to Chris Coey and Juan Pablo Vielma for stimulating discussions on topics discussed in this paper.

## References

Arrow, KJ (1965) Aspects of the theory of risk-bearing. Yrjö Jahnssonin Säätiö

Balas E, Ceria S, Cornuéjols G (1993) A lift-and-project cutting plane algorithm for mixed 0–1 programs. Mathematical programming **58**:295–324.

Beasley JE (1990) Or-library: distributing test problems by electronic mail. The Journal of the Operational Research Society **41**(11):1069–1072.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) Robust optimization (Princeton University Press).

Ben-Tal A, Nemirovski A (2001) Lectures on modern convex optimization, volume **2** (SIAM).

Bertsekas DP (1999) Nonlinear programming (Athena scientific Belmont).

Bertsimas D, Darnell C, Soucy R (1999) Portfolio construction through mixed-integer programming at Grantham, Mayo, van Otterloo and company. Interfaces **29**(1):49–66.

Bertsimas D, King A, Mazumder R (2016) Best subset selection via a modern optimization lens. The annals of statistics. **44**(2):813–852.

Bertsimas D, Pauphilet J, Van Parys B (2017) Sparse classification and phase transitions: a discrete optimization perspective, Journal of Machine Learning Research, Under Review.

Bertsimas D, Popescu I (2005) Optimal inequalities in probability theory: A convex optimization approach. SIAM Journal on Optimization. **15**(3):780–804.

Bertsimas D, Shioda R (2009) Algorithm for cardinality-constrained quadratic optimization. Computational optimization and applications. **43**(1):1–22.

Bertsimas D, Van Parys B (2016) Sparse high dimensional regression: Exact scalable algorithms and phase transitions, The annals of statistics., Under Review.

Bezanson J, Edelman A, Karpinski S, Shah VB (2017) Julia: A fresh approach to numerical computing. SIAM Review **59**(1):65–98.

Bienstock D (1996) Computational study of a family of mixed-integer quadratic programming problems. Mathematical Programming **74**(2):121–140.

Bonami P, Lejeune MA (2009) An exact solution approach for portfolio optimization problems under stochastic and integer constraints. Operations Research **57**(3):650–670.

Brodie J, Daubechies I, De Mol C, Giannone D, Loris I (2009) Sparse and stable Markowitz portfolios. Proceedings of the National Academy of Sciences **106**(30):12267–12272.

Chang TJ, Meade N, Beasley JE, Sharaiha YM (2000) Heuristics for cardinality constrained portfolio optimisation. Computers & Operations Research **27**(13):1271–1302.

CPLEX II (2018) V12. 8: User's manual for CPLEX. International Business Machines Corporation .

Cui X, Zheng X, Zhu S, Sun X (2013) Convex relaxations and miqcqp reformulations for a class of cardinality-constrained portfolio selection problems. Journal of Global Optimization **56**(4):1409–1423.

Danna E, Rothberg E, Le Pape C (2005) Exploring relaxation induced neighborhoods to improve mip solutions. Mathematical Programming. **102**(1):71–90.

DeMiguel V, Garlappi L, Uppal R (2007) Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? The review of Financial studies **22**(5):1915–1953.

DeMiguel V, Garlappi L, Nogales FJ, Uppal R (2009) A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. Management Science **55**(5):798–812.

Dunning I, Huchette J, Lubin M (2017) JuMP: A modeling language for mathematical optimization. SIAM Review **59**(2):295–320.

Ehrgott M (2005) Multicriteria optimization, volume **491** of Lecture Notes in Economics and Mathematical Systems (Springer).

Fletcher R, Leyffer S (1994) Solving mixed integer nonlinear programs by outer approximation. Mathematical Programming. **66**(1):327–349.

Frangioni A, Gentile C (2006) Perspective cuts for a class of convex 0–1 mixed integer programs. Mathematical Programming. **106**:225–236.

Gao J, Li D (2013) Optimal cardinality constrained portfolio selection. Operations Research **61**(3):745–761.

Gurobi Optimization (2018) Gurobi optimizer reference manual; 2018. www.gurobi.com .

Kelley JE Jr (1960) The cutting-plane method for solving convex programs. Journal of the society for Industrial and Applied Mathematics. **8**(4):703–712.

Lobo MS, Vandenberghe L, Boyd S, Lebret H (1998) Applications of second-order cone programming. *Linear Algebra and its Applications* **284**(1-3):193–228.

Lodi A (2013) The heuristic (dark) side of MIP solvers. Hybrid metaheuristics 273–284

Markowitz H (1952) Portfolio selection. The Journal of Finance. **7**(1):77–91.

Koç A, Morton DP (2014) Prioritization via stochastic optimization. *Management Science* **61**(3):586–603.

Mosek A (2010) The mosek optimization software. Online at http://www. mosek. com **54**(2-1):5.

Nesterov Y (2013) Introductory lectures on convex optimization: A basic course, volume **87** (Springer).

Olivares-Nadal AV, DeMiguel V (2018) A robust perspective on transaction costs in portfolio optimization. *Operations Research* Articles in Advance.

Pflug GC, Pichler A, Wozabal D (2012) The 1/n investment strategy is optimal under high model ambiguity. Journal of Banking & Finance **36**(2):410–417.

Pilanci M, Wainwright MJ, El Ghaoui L (2015) Sparse learning via boolean relaxations. Mathematical Programming. **151**(1):63–87.

Pratt JW (1975) Risk aversion in the small and in the large. *Stochastic Optimization Models in Finance*, 115–130 (Elsevier).

Ryan JA, Ulrich JM, Thielen W, Teetor P, Bronder S (2018) quantmod: Quantitative financial modelling framework. R package version 0.4-0. http://www. quantmod. com

Shioda R (2003) Integer optimization in data mining. PhD Thesis, Massachusetts Institute of Technology.

Tikhonov A (1963) Solution of incorrectly formulated problems and the regularization method. Soviet Math. Dokl. **5**:1035–1038.

Vielma JP, Ahmed S, Nemhauser GL (2008) A lifted linear programming branch-and-bound algorithm for mixed-integer conic quadratic programs. INFORMS Journal on Computing. **20**(3):438–450.

Vielma JP, Dunning I, Huchette J, Lubin M (2017) Extended formulations in mixed integer conic quadratic programming. Mathematical Programming Computation. **9**(3):369–418.

Wainwright MJ (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using $l_1$-constrained quadratic programming (lasso). IEEE Transactions on Information Theory **55**(5):2183–2202.

Zakeri G, Craigie D, Philpott A, Todd M (2014) Optimization of demand response through peak shaving. Operations Research Letters **42**(1):97–101.