Autonomous Sparse Mean-CVaR Portfolio Optimization

Yizun Lin [1] **Yangyu Zhang** [2] **Zhao-Rong Lai** [1] **Cheng Li** [1]

Abstract

The ℓ_0-constrained mean-CVaR model poses a significant challenge due to its NP-hard nature, typically tackled through combinatorial methods characterized by high computational demands. From a markedly different perspective, we propose an innovative autonomous sparse mean-CVaR portfolio model, capable of approximating the original ℓ_0-constrained mean-CVaR model with arbitrary accuracy. The core idea is to convert the ℓ_0 constraint into an indicator function and subsequently handle it through a tailed approximation. We then propose a proximal alternating linearized minimization algorithm, coupled with a nested fixed-point proximity algorithm (both convergent), to iteratively solve the model. Autonomy in sparsity refers to retaining a significant portion of assets within the selected asset pool during adjustments in pool size. Consequently, our framework offers a theoretically guaranteed approximation of the ℓ_0-constrained mean-CVaR model, improving computational efficiency while providing a robust asset selection scheme.

1. Introduction

Value-at-risk (VaR, Jorion 1997) is a widely-used downside-risk metric in finance that assesses the extreme percentage loss in a disadvantageous situation. Since it fails to satisfy the subadditivity property (Artzner et al., 1999), it is improved to its tail statistic named conditional value-at-risk (CVaR, Rockafellar & Uryasev 2000). CVaR not only has the good theoretical property of coherence (Artzner et al., 1999; Gotoh & Takeda, 2011), but also accords with dual stochastic dominance (Ogryczak & Ruszczyński, 2002). Moreover, CVaR is more tractable than VaR from the perspective of optimization. It can serve as the risk metric in Markowitz's mean-risk criterion (Markowitz, 1952) which balances the expected return and the risk of a portfolio. The original mean-CVaR model with linear constraints can be efficiently solved by linear programming (Rockafellar & Uryasev, 2000).

In practical portfolio optimization (PO), it is desirable to constrain the scale of selected assets while keeping an appropriate investing performance, in order to reduce transaction cost and managerial burden. This can be done via the managerial approach, such as the endowment model (Dimmock et al., 2023), market crashes (Liu & Loewenstein, 2013), and the revenue-driven resource allocation (Chao et al., 2009). However, these methods still demand professional knowledge in management and finance. Recently, the arising development in machine learning methods for management science provides an alternative way for PO. For example, (Ban et al., 2018) propose the performance-based regularization with a data-driven strategy to improve the out-of-sample performance in PO. (Brodie et al., 2009) adopt the ℓ_1 regularization to construct a sparse and stable Markowitz portfolio (SSMP). (Lai et al., 2018) and (Luo et al., 2020) propose short-term sparse portfolio optimization models with ℓ_1 and ℓ_0 regularizations, respectively. (Lai et al., 2020) propose a sparse structure for the covariance estimation based on the spectral decomposition.

Among all the sparsity approaches, the ℓ_0 constraint can exactly control the number of selected assets. However, the ℓ_0 constraint generally leads to an NP-hard problem that has a very high computational complexity. A general approach to exactly solve the ℓ_0-constrained problems is the combinatorial approach, such as the mixed-integer optimization (Günlük & Linderoth, 2010) and the branch-and-bound procedure (Kobayashi et al., 2021). For example, (Bertsimas & Cory-Wright, 2022) propose a cutting-plane approach with some heuristics to exactly solve an ℓ_0-constrained mean-variance model, which saves some computational time according to the experimental results. (Kobayashi et al., 2021) further adopt this approach to exactly solve an ℓ_0-constrained mean-CVaR model. It is essentially a branch-and-bound algorithm: for each visit of a node (a combinatorial case), it conducts one ordinary mean-CVaR optimization. Since its experimental results show

[1]Department of Mathematics, College of Information Science and Technology, Jinan University, Guangzhou, China [2]Jinan University-University of Birmingham Joint Institute, Jinan University, Guangzhou, China. Correspondence to: Zhao-Rong Lai <laizhr@jnu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

that the number of visited nodes ranges from about 1,300 to more than 77,000, the computational complexity is still high in practical portfolio management. Moreover, (Kobayashi et al., 2021) requires the commercial optimizer Gurobi[1] to deploy this cutting-plane approach. We try to run the program, but it prompts an error that the models are too large in our experiments. In summary, the existing exactly-solving approaches have high computational complexity and rely on the external commercial optimizer, thus economic and practical approaches still need to be developed.

Instead of exactly solving the ℓ_0-constrained mean-CVaR model, we propose an autonomous sparse mean-CVaR (ASMCVaR) PO model that can approximate it to arbitrary accuracy. The core idea is to convert the ℓ_0 constraint into an indicator function and approximate it to arbitrary accuracy through a tailed approximation approach (Definition 3.1). By this way, we can replace the nonconvex and complex geometric structure of ℓ_0 constraint by a simple and more tractable tailed approximation. To solve the ASMCVaR model, we develop a Proximal Alternating Linearized Minimization Algorithm (PALM) with a nested Fixed-Point Proximity Algorithm (FPPA) that can handle all the constraints and guarantee convergence. It saves much computational complexity and keeps a certain accuracy at the same time. ASMCVaR can keep a large proportion of assets in the selected asset pool when adjusting the pool size, which reduces managerial burden and delivers convenience. Our main contributions can be summarized as follows.

1. We propose an ASMCVaR model and prove that it can approximate the ℓ_0-constrained mean-CVaR model to arbitrary accuracy based on item 2.
2. We propose a tailed approximation for the indicator function of the ℓ_0 constraint. We also find that this approximation can be arbitrarily accurate as the positive approximation parameter tends to 0. See Definition 3.1 and Figure 1 for more detail.
3. We propose a convergent PALM algorithm with a nested FPPA to iteratively solve ASMCVaR. It does not require any external or commercial optimizers.
4. ASMCVaR can keep a large proportion of assets in the selected asset pool when adjusting the pool size, thus the selected assets are stable. It is beneficial to practical portfolio management when the manager needs to adjust the asset pool size.

2. Preliminaries and Related Works

We present some preliminaries on CVaR, and then introduce some approximate or exact solutions to the ℓ_0-constrained mean-CVaR model.

[1] https://www.gurobi.com/

2.1. VaR and CVaR

VaR measures the percentage loss of portfolio return in a bad scenario and with a certain confidence level c:

$$\Pr\{\mathrm{VaR} \leqslant -r_s\} = 1 - c,$$

where $\Pr\{\cdot\}$ denotes the probability of an event and r_s denotes the rate of return of a portfolio (representing some kind of investing strategy). Specifically, if $\mathrm{VaR}(c = 95\%) = 0.1$, then this investing strategy may suffer a loss greater than 10% with a probability of 5%.

CVaR further averages the values greater than the VaR at a given confidence level c:

$$\mathrm{CVaR} := \min_{\tau\in\mathbb{R}} \left\{ \tau + \frac{1}{1-c}\mathbb{E}\left(\max\{-r_s - \tau, 0\}\right) \right\},$$

where $\mathbb{E}(\cdot)$ denotes the expectation of a random variable.

2.2. Original Mean-CVaR Model

In a real-world investing environment, the CVaR can be formulated as a linear program. Let $\boldsymbol{R} := [\boldsymbol{r}_1^\top; \cdots; \boldsymbol{r}_T^\top] \in \mathbb{R}^{T\times N}$ be the sample return matrix of N assets and T observations gathered from a financial market. Accordingly, a portfolio (investing strategy) can be seen as an N-dimensional vector $\boldsymbol{w}$. Then the sample returns of this portfolio are $\{\boldsymbol{r}_t^\top \boldsymbol{w}\}_{t=1}^T$. Using the sample average estimator for CVaR yields

$$\min_{\boldsymbol{w}\in\mathbb{R}^N,\ \tau\in\mathbb{R},\ \boldsymbol{z}\in\mathbb{R}^T} \left\{ \tau + \frac{1}{(1-c)T}\mathbf{1}_T^\top \boldsymbol{z} \right\} \tag{1}$$

$$\text{s.t.} \quad \boldsymbol{z} \geqslant -\boldsymbol{R}\boldsymbol{w} - \tau\mathbf{1}_T, \quad \boldsymbol{z} \geqslant \mathbf{0}_T, \tag{2}$$

$$\boldsymbol{w} \in \Delta := \left\{ \boldsymbol{w} \in \mathbb{R}^N : \boldsymbol{w} \geqslant \mathbf{0}_N, \boldsymbol{w}^\top \mathbf{1}_N = 1 \right\}, \tag{3}$$

$$(\boldsymbol{w}^\top \hat{\boldsymbol{\mu}} = \rho), \tag{4}$$

where $\boldsymbol{z}$ is the auxiliary variable to deal with the positive part in the expectation. $\mathbf{1}_T$ or $\mathbf{0}_T$ is a vector of 1 or 0 with dimensionality shown in the subscript, respectively. The inequality for vectors dominates each dimension. In (2), inequality constraints for $\boldsymbol{z}$ can be used instead of equality constraints because $\boldsymbol{z}$ is greater than $\mathbf{0}_T$ and the objective is a minimization. (3) is the simplex constraint that consists of the no-short-selling constraint and the self-financing constraint, respectively. Such constraints ensure feasibility of the portfolio in real-world investment. (4) is an optional constraint for the expected portfolio return, where $\hat{\boldsymbol{\mu}} = \frac{1}{T}\boldsymbol{R}^\top \mathbf{1}_T$ is the sample asset return vector. (1)∼(4) constitute the original mean-CVaR model (Rockafellar & Uryasev, 2000). Since all the constraints are convex and linear, this model can be efficiently solved by the simplex method.

2.3. Mean-CVaR Model with ℓ_1 Regularization

A widely adopted approach for sparsity is to impose ℓ_1 regularization. This method is favored for its compatibil-

ity with widely recognized solving techniques, such as the linear programming, the Lasso (Tibshirani, 1996) and the LARS (Efron et al., 2004) algorithms. (Cheng & Gao, 2015) propose a mean-CVaR model with a re-weighted ℓ_1 regularization as follows:

$$\min_{\boldsymbol{w}\in\mathbb{R}^N,\tau\in\mathbb{R},\boldsymbol{z}\in\mathbb{R}^T}\left\{\tau+\frac{1}{(1-c)T}\mathbf{1}_T^\top\boldsymbol{z}+\boldsymbol{\lambda}^\top\boldsymbol{w}\right\} \quad (5)$$
$$\text{s.t. } \boldsymbol{z}\geqslant -\boldsymbol{R}\boldsymbol{w}-\tau\mathbf{1}_T, \boldsymbol{z}\geqslant \mathbf{0}_T, \boldsymbol{w}\in\Delta, (\boldsymbol{w}^\top\hat{\boldsymbol{\mu}}=\rho),$$

where $\boldsymbol{\lambda}\in\{\boldsymbol{x}\in\mathbb{R}^N : \boldsymbol{x}\geqslant \mathbf{0}_N\}$ is the weight vector for the ℓ_1 regularization. Since $\boldsymbol{w}\geqslant\mathbf{0}_N$, the re-weighted ℓ_1 regularization in (5) can be simplified: $\sum_{i=1}^N\lambda_i|w_i| = \sum_{i=1}^N\lambda_i w_i = \boldsymbol{\lambda}^\top\boldsymbol{w}$. This formulation can also be efficiently solved by linear programming. In each outer-iteration, the algorithm updates $\boldsymbol{\lambda}$ by some rules regarding $\boldsymbol{w}$, for example, $\lambda_i^{k+1}=1/(|w_i^k|+\epsilon)$. Then in the inner-iteration, (5) is solved with fixed $\boldsymbol{\lambda}^{k+1}$ to obtain $\boldsymbol{w}^{k+1}$. This approach enjoys a low computational cost, but it cannot exactly control the number of selected assets. One can just empirically tune $\boldsymbol{\lambda}$ to get a rough number of selected assets. Besides, it is also intractable to assess the accuracy of approximation of ℓ_1 regularization to the ℓ_0 constraint.

2.4. ℓ_0-constrained Mean-CVaR Model

To impose exact sparsity on the portfolio, the ℓ_0 constraint should be used:

$$\min_{\boldsymbol{w}\in\mathbb{R}^N,\ \tau\in\mathbb{R},\ \boldsymbol{z}\in\mathbb{R}^T}\left\{\tau+\frac{1}{(1-c)T}\mathbf{1}_T^\top\boldsymbol{z}\right\} \quad (6)$$
$$\text{s.t. } \boldsymbol{z}\geqslant -\boldsymbol{R}\boldsymbol{w}-\tau\mathbf{1}_T,\ \boldsymbol{z}\geqslant\mathbf{0}_T,\ \boldsymbol{w}\in\Delta,\ (\boldsymbol{w}^\top\hat{\boldsymbol{\mu}}=\rho),$$
$$\|\boldsymbol{w}\|_0\leqslant m, \quad (7)$$

where Δ is defined in (3), and $\|\boldsymbol{w}\|_0$ represents the number of nonzero components in $\boldsymbol{w}$. The ℓ_0 constraint in (7) is also called the m-sparse constraint, since m controls the number of selected assets. The above ℓ_0-constrained mean-CVaR problem is NP-hard. To solve it by the combinatorial approach, (Kobayashi et al., 2021) introduce an auxiliary variable $\boldsymbol{y}\in\mathcal{Z}_N^m := \{\boldsymbol{y}\in\{0,1\}^N : \mathbf{1}_N^\top\boldsymbol{y}\leqslant m\}$, then (7) is equivalent to the logical constraint

$$\boldsymbol{y}\in\mathcal{Z}_N^m, \quad y_i=0\Rightarrow w_i=0. \quad (8)$$

Let $\boldsymbol{Y}:=\mathrm{diag}(\boldsymbol{y})$ be the diagonalized matrix for $\boldsymbol{y}$. Then model (6) can be reformulated as a bilevel optimization problem

$$\text{(upper-level)} \qquad \min_{\boldsymbol{y}\in\mathcal{Z}_N^m} f(\boldsymbol{y}), \quad (9)$$

(lower-level)

$$f(\boldsymbol{y})=\min_{\boldsymbol{w}\in\mathbb{R}^N,\tau\in\mathbb{R},\boldsymbol{z}\in\mathbb{R}^T}\left\{\frac{1}{2\gamma}\boldsymbol{w}^\top\boldsymbol{w}+\tau+\frac{1}{(1-c)T}\mathbf{1}_T^\top\boldsymbol{z}\right\} \quad (10)$$

$$\text{s.t. } \boldsymbol{z}\geqslant -\boldsymbol{R}\boldsymbol{Y}\boldsymbol{w}-\tau\mathbf{1}_T, \boldsymbol{z}\geqslant\mathbf{0}_T, \boldsymbol{Y}\boldsymbol{w}\in\Delta, (\boldsymbol{w}^\top\boldsymbol{Y}\hat{\boldsymbol{\mu}}=\rho), \quad (11)$$

where the logical constraint (8) is embedded in $\boldsymbol{Y}\boldsymbol{w}$. Note that $\boldsymbol{w}=\boldsymbol{Y}\boldsymbol{w}$ after minimization (10), thus $(\boldsymbol{Y}\boldsymbol{w})^\top\boldsymbol{Y}\boldsymbol{w}$ is replaced by $\boldsymbol{w}^\top\boldsymbol{w}$ (Bertsimas & Cory-Wright, 2022). As $\gamma\to+\infty$, this bilevel optimization can approximate the ℓ_0-constrained mean-CVaR optimization to arbitrary accuracy.

In the upper-level optimization, (Kobayashi et al., 2021) solve the following problem

$$\min_{(\boldsymbol{y},\theta)\in\mathcal{Z}_N^m\times\mathbb{R}}\theta \quad \text{s.t.} \quad (\boldsymbol{y},\theta)\in\mathcal{F}^k\subseteq\mathcal{Z}_N^m\times\mathbb{R} \quad (12)$$

to obtain the node $(\boldsymbol{y}^k,\theta^k)$ of the k-th visit, where $\mathcal{F}^k$ denotes the k-th feasible region for $(\boldsymbol{y},\theta)$. θ indicates any possible value for $f(\boldsymbol{y})$ in the current iteration, and its infimum provides a lower bound for $f(\boldsymbol{y})$. To initialize, one can just solve a continuously-relaxed and lower-bounded version of (6)∼(7) to obtain θ^1 for all $\boldsymbol{y}\in\mathcal{Z}_N^m$ and pick up any $\boldsymbol{y}^1\in\mathcal{Z}_N^m$. In the lower-level optimization, (10)∼(11) is a convex quadratic optimization for which efficient algorithms exist. If it is infeasible, then $\mathcal{F}^{k+1}$ is obtained by removing $(\boldsymbol{y}^k,\theta^k)$ from $\mathcal{F}^k$; or else it can be solved to obtain $f(\boldsymbol{y}^k)$. Since $f(\boldsymbol{y})$ is convex in $\boldsymbol{y}\in[0,1]^N$, the set

$$\{(\boldsymbol{y},\theta)\in\mathcal{Z}_N^m\times\mathbb{R} : \theta\geqslant f(\boldsymbol{y}^k)+\boldsymbol{g}_{\boldsymbol{y}^k}^\top(\boldsymbol{y}-\boldsymbol{y}^k)\} \quad (13)$$

provides new lower bounds for all the unvisited cases $\boldsymbol{y}\in\mathcal{Z}_N^m$, where $\boldsymbol{g}_{\boldsymbol{y}^k}$ is a subgradient of f at $\boldsymbol{y}^k$. Intersecting (13) and $\mathcal{F}^k$ yields $\mathcal{F}^{k+1}$. Either feasible or infeasible, the lower-level optimization narrows down the searching region from $\mathcal{F}^k$ to $\mathcal{F}^{k+1}$. Then in the next visit (iteration), (12) is solved in $\mathcal{F}^{k+1}$. During this procedure, $LB^k := \theta^k$ serves as the current uniform lower bound for all the unvisited cases, while $UB^k := \min_{1\leqslant j\leqslant k} f(\boldsymbol{y}^j)$ serves as the current uniform upper bound for all the visited cases. If $(UB^k - LB^k < \epsilon)$ for a sufficiently small positive ϵ, then $\{\boldsymbol{y}^j : 1\leqslant j\leqslant k, f(\boldsymbol{y}^j)=UB^k\}$ can be considered as the ϵ-optimal solution set. In fact, this is a branch-and-bound strategy.

Though the above bilevel optimization takes the exact solving objective as the initial intention, it involves two approximations: one is the term $\frac{1}{2\gamma}\boldsymbol{w}^\top\boldsymbol{w}$ in (10), the other is the ϵ optimality of the gap $(UB^k - LB^k)$. Moreover, this branch-and-bound approach may visit tens of thousands of combinatorial cases to achieve satisfactory optimality. In addition, external commercial optimizers are also required to deploy the whole algorithm. The program is unable to run in our experiments because the models are too large. To develop economic and practical methods, we propose to approximate the ℓ_0-constrained mean-CVaR model from a very different perspective based on a tailed approximation.

3. Autonomous Sparse Mean-CVaR Portfolio Optimization

In this section, we propose an Autonomous Sparse Mean-CVaR (ASMCVaR) model and develop a Proximal Alter-

nating Linearized Minimization (PALM) algorithm to solve this model.

3.1. Autonomous Sparse Mean-CVaR Model

Before introducing the ASMCVaR model, we revise the ℓ_0-constrained Mean-CVaR model and reformulate it as a nonconvex three-term optimization model. Note that there is an equality constraint $\boldsymbol{w}^\top \boldsymbol{\mu} = \rho$ in the Mean-CVaR model (6). In general, the expected return level ρ is generally set to a positive number. However, this constraint is not always feasible not only in reality but also in theory. The satisfaction of a tight equality constraint for the return level in PO with an ever-changing financial market is difficult. To avoid this problem, we use a regularization term $(\boldsymbol{w}^\top \hat{\boldsymbol{\mu}} - \rho)^2$ in the objective function instead of using the equality constraint. In this way, the portfolio return $\boldsymbol{w}^\top \hat{\boldsymbol{\mu}}$ is balanced between the expected level ρ and the risk metric CVaR. Then the revised ℓ_0-constrained Mean-CVaR model is given by

$$\begin{aligned} &\min_{\boldsymbol{w}\in\mathbb{R}^N,\ \tau\in\mathbb{R},\ \boldsymbol{z}\in\mathbb{R}^T} \left\{\tau + \frac{1}{(1-c)T}\mathbf{1}_T^\top \boldsymbol{z} + \lambda(\boldsymbol{w}^\top \boldsymbol{\mu} - \rho)^2\right\} \\ &\text{s.t. } \boldsymbol{z} \geqslant -\boldsymbol{R}\boldsymbol{w} - \tau \mathbf{1}_T,\ \boldsymbol{z} \geqslant \mathbf{0}_T,\ \boldsymbol{w} \in \Delta,\ \|\boldsymbol{w}\|_0 \leqslant m. \end{aligned} \tag{14}$$

For notation simplicity, we define $N_1 := N + 1 + T$, $N_2 := 2T + N + 2$. We then rewrite model (14) as a more compact form. To this end, we let $\boldsymbol{v} := (\boldsymbol{w}^\top, \tau, \boldsymbol{z}^\top)^\top \in \mathbb{R}^{N_1}$,

$$\boldsymbol{h}_1 := \left(\mathbf{0}_N^\top, 1, \frac{1}{(1-c)T}\mathbf{1}_T^\top\right)^\top,\ \boldsymbol{h}_2 := \left(\boldsymbol{\mu}^\top, \mathbf{0}_{1+T}^\top\right)^\top, \tag{15}$$

$$\tilde{\boldsymbol{Q}} := \begin{pmatrix} \boldsymbol{R} & \mathbf{1}_T & \boldsymbol{I}_T \\ \mathbf{0}_{T\times N} & \mathbf{0}_T & \boldsymbol{I}_T \end{pmatrix} \in \mathbb{R}^{2T\times N_1},$$

$$\widetilde{\boldsymbol{I}} := \begin{pmatrix} \boldsymbol{I}_N & \mathbf{0}_{N\times(T+1)} \end{pmatrix} \in \mathbb{R}^{N\times N_1},$$

and define

$$\boldsymbol{Q} := \begin{pmatrix} & \tilde{\boldsymbol{Q}} & \\ \boldsymbol{I}_N & \mathbf{0}_{N\times(T+1)} \\ \mathbf{1}_N^\top & \mathbf{0}_{T+1}^\top \\ -\mathbf{1}_N^\top & \mathbf{0}_{T+1}^\top \end{pmatrix} \in \mathbb{R}^{N_2\times N_1},$$

$$\boldsymbol{q} := \begin{pmatrix} \mathbf{0}_{2T+N} \\ 1 \\ -1 \end{pmatrix} \in \mathbb{R}^{N_2}. \tag{16}$$

Then model (14) can be rewritten as the following model with respect to (w.r.t.) $\boldsymbol{v}$:

$$\begin{aligned} &\min_{\boldsymbol{v}\in\mathbb{R}^{N_1}} \left\{\boldsymbol{h}_1^\top \boldsymbol{v} + \lambda(\boldsymbol{h}_2^\top \boldsymbol{v} - \rho)^2\right\} \\ &\text{s.t. } \boldsymbol{Q}\boldsymbol{v} \geqslant \boldsymbol{q},\ \ \|\tilde{\boldsymbol{I}}\boldsymbol{v}\|_0 \leqslant m. \end{aligned} \tag{17}$$

Model (17) comprises two constraints. In fact, it can be reformulated as an equivalent three-term unconstrained minimization model. For this purpose, we define two indicator functions

$$\iota_{\boldsymbol{q}}(\boldsymbol{u}) := \begin{cases} 0, & \text{if } \boldsymbol{u} \geqslant \boldsymbol{q}; \\ +\infty, & \text{otherwise}, \end{cases}$$

$$\iota_m(\boldsymbol{w}) := \begin{cases} 0, & \text{if } \|\boldsymbol{w}\|_0 \leqslant m; \\ +\infty, & \text{otherwise}, \end{cases}$$

for $\boldsymbol{u} \in \mathbb{R}^{N_2}$ and $\boldsymbol{w} \in \mathbb{R}^N$, respectively; and define

$$f(\boldsymbol{v}) := \boldsymbol{h}_1^\top \boldsymbol{v} + \lambda(\boldsymbol{h}_2^\top \boldsymbol{v} - \rho)^2,\ \ \boldsymbol{v} \in \mathbb{R}^{N_1}. \tag{18}$$

Then model (17) is equivalent to the following model:

$$\min_{\boldsymbol{v}\in\mathbb{R}^{N_1}} \left\{\Phi(\boldsymbol{v}) := f(\boldsymbol{v}) + \iota_{\boldsymbol{q}}(\boldsymbol{Q}\boldsymbol{v}) + \iota_m(\widetilde{\boldsymbol{I}}\boldsymbol{v})\right\}. \tag{19}$$

Directly solving model (19) is very tough due to the nondifferentiability of $\iota_{\boldsymbol{q}}$ and the nonconvexity of ι_m. Instead, we introduce a surrogate model of model (19) and develop an efficient algorithm to solve the surrogate model. To this end, we introduce a function to approximate the exact m-sparse function ι_m. For vector $\boldsymbol{w} \in \mathbb{R}^N$, let $\{j_1, j_2, \ldots, j_N\}$ be any rearrangement of $\{1, 2, \cdots, N\}$ such that $|w_{j_1}| \geqslant |w_{j_2}| \geqslant \cdots \geqslant |w_{j_N}|$. Then we call $\{j_1, j_2, \cdots, j_m\}$ an m-largest absolute value (m-LAV) index set of $\boldsymbol{w}$. Note that the m-LAV index sets of $\boldsymbol{w}$ may not be unique since the components of $\boldsymbol{w}$ may not be distinct. Throughout this paper, for $\boldsymbol{w} \in \mathbb{R}^N$, we denote by $J_{\boldsymbol{w}}^m$ an m-LAV index set of $\boldsymbol{w}$. By smoothing the giant leap of ι_m between $\|\boldsymbol{w}\|_0 \leqslant m$ and $\|\boldsymbol{w}\|_0 > m$, we construct the following tailed approximation.

Definition 3.1 (Tailed Approximation of ℓ_0 Constraint)**.** The tailed approximation for the m-sparse indicator function is defined by

$$\tilde{\iota}_{m,\gamma}(\boldsymbol{w}) = \begin{cases} 0, & \text{if } \|\boldsymbol{w}\|_0 \leqslant m; \\ \frac{1}{2\gamma} \sum\limits_{j\in\mathbb{N}_N\setminus J_{\boldsymbol{w}}^m} w_j^2, & \text{if } \|\boldsymbol{w}\|_0 > m. \end{cases} \tag{20}$$

where $\gamma > 0$ is the approximation parameter.

When $\|\boldsymbol{w}\|_0 > m$, the tail term $\frac{1}{2\gamma} \sum\limits_{j\in\mathbb{N}_N\setminus J_{\boldsymbol{w}}^m} w_j^2 > 0$, then $\frac{1}{2\gamma} \sum\limits_{j\in\mathbb{N}_N\setminus J_{\boldsymbol{w}}^m} w_j^2 \nearrow +\infty$ as $\gamma \searrow 0$. Hence for any fixed $\boldsymbol{w}$, either $\tilde{\iota}_{m,\gamma}(\boldsymbol{w}) = \iota_m(\boldsymbol{w})$ ($\|\boldsymbol{w}\|_0 \leqslant m$) or $\tilde{\iota}_{m,\gamma}(\boldsymbol{w}) \nearrow \iota_m(\boldsymbol{w})$ as $\gamma \searrow 0$ ($\|\boldsymbol{w}\|_0 > m$). Figure 1 shows a diagram for $\tilde{\iota}_{m,\gamma}(\boldsymbol{w})$ with $N = 10$ and $m = 5$.

Now the proposed ASMCVaR model is given by

$$\min_{\boldsymbol{v}\in\mathbb{R}^{N_1}} \left\{\Psi(\boldsymbol{v}) := f(\boldsymbol{v}) + \iota_{\boldsymbol{q}}(\boldsymbol{Q}\boldsymbol{v}) + \tilde{\iota}_{m,\gamma}(\widetilde{\boldsymbol{I}}\boldsymbol{v})\right\}. \tag{21}$$

We then establish the relationship between the solutions of model (19) and the ASMCVaR model. For this purpose, we need the following technical lemma, whose proof is given in Supplementary A.1.

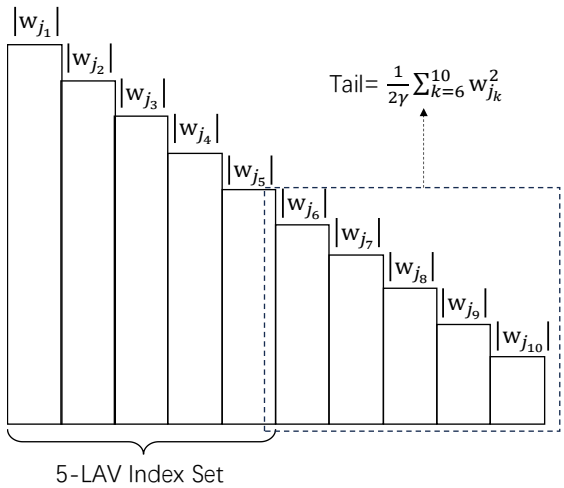

Figure 1. Diagram for tailed approximation of ℓ_0 constraint: $\tilde{\iota}_{m,\gamma}(\boldsymbol{w})$ with $N = 10$ and $m = 5$. As $\gamma \searrow 0$, $\frac{1}{2\gamma}\sum_{k=6}^{10} w_{j_k}^2 \nearrow +\infty$ and $\tilde{\iota}_{m,\gamma}(\boldsymbol{w}) \nearrow \iota_m(\boldsymbol{w})$ for any fixed $\boldsymbol{w}$ ($\|\boldsymbol{w}\|_0 > 5$).

Lemma 3.2. *Let $\boldsymbol{v}^*$ be a solution of model (21) and $\boldsymbol{w}^* := \tilde{\boldsymbol{I}}\boldsymbol{v}^*$. Then there exists a constant $\tilde{L} > 0$ such that*

$$w_j^* \leqslant \sqrt{2\tilde{L}\gamma}, \ \text{for all } j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}^*}^m. \tag{22}$$

The following theorem indicates that the optimal value of f in (21) can approximate that in (19) to arbitrary accuracy by letting $\gamma \searrow 0$.

Theorem 3.3. *Let $\boldsymbol{v}^*$ and $\boldsymbol{v}^{**}$ be solutions of models (21) and (19), respectively. Then there exists a constant $\tilde{L} > 0$ such that*

$$0 \leqslant f(\boldsymbol{v}^{**}) - f(\boldsymbol{v}^*) \leqslant \sqrt{2\tilde{L}^3\gamma}. \tag{23}$$

The proof is given in Supplementary A.2. To solve model (21), we further rewrite it as an optimization model w.r.t. two variables $\boldsymbol{v}$ and $\boldsymbol{y}$ as follows:

$$\min_{\boldsymbol{v}\in\mathbb{R}^{N_1},\ \boldsymbol{y}\in\mathbb{R}^N} \left\{G(\boldsymbol{v},\boldsymbol{y}) := H(\boldsymbol{v},\boldsymbol{y}) + \iota_{\boldsymbol{q}}(\boldsymbol{Q}\boldsymbol{v}) + \iota_m(\boldsymbol{y})\right\}, \tag{24}$$

where

$$H(\boldsymbol{v},\boldsymbol{y}) := f(\boldsymbol{v}) + \frac{1}{2\gamma}\left\|\tilde{\boldsymbol{I}}\boldsymbol{v} - \boldsymbol{y}\right\|_2^2, \ \ \gamma > 0. \tag{25}$$

This model can be directly solved by the proximal alternating linearized minimization algorithm (Attouch et al., 2013; Bolte et al., 2014). We shall then prove the equivalence of the solutions of models (21) and (24). To this end, we need the following two lemmas. We let

$$\Omega_m := \{\boldsymbol{w} \in \mathbb{R}^N : \|\boldsymbol{w}\|_0 \leqslant m\}, \tag{26}$$

and define the thresholding operator $\mathcal{S}_m$ w.r.t. the m largest components by

$$[\mathcal{S}_m(\boldsymbol{w})]_j := \begin{cases} w_j, & \text{if } j \in J_{\boldsymbol{w}}^m; \\ 0, & \text{if } j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}}^m, \end{cases} \tag{27}$$

for $\boldsymbol{w} \in \mathbb{R}^N$. Note that $\mathcal{S}_m(\boldsymbol{w})$ may not be unique and it depends on a corresponding m-LAV index set of $\boldsymbol{w}$. Throughout this paper, the equality $\boldsymbol{x} = \mathcal{S}_m(\boldsymbol{w})$ represents $\boldsymbol{x} \in \mathcal{S}_m(\boldsymbol{w})$.

Lemma 3.4. *If a pair $(\boldsymbol{v}^*, \boldsymbol{y}^*)$ is a solution of model (24), then $\boldsymbol{y}^* = \mathcal{S}_m(\tilde{\boldsymbol{I}}\boldsymbol{v}^*)$.*

Lemma 3.5. *Let G be defined in (24), $\boldsymbol{v} \in \mathbb{R}^{N_1}$. If $\tilde{\boldsymbol{y}} = \mathcal{S}_m(\tilde{\boldsymbol{I}}\boldsymbol{v})$, then*

$$G(\boldsymbol{v},\tilde{\boldsymbol{y}}) \leqslant G(\boldsymbol{v},\boldsymbol{y}), \ \forall \boldsymbol{y} \in \mathbb{R}^N. \tag{28}$$

The proof of Lemma 3.4 and Lemma 3.5 are given in Supplementary A.3 and A.4, respectively. According to these two lemmas, we have the following theorem, whose proof is provided in Supplementary A.5.

Theorem 3.6. *Let functions Ψ and G be defined by (21) and (24), respectively. Then*

$$\Psi(\boldsymbol{v}) = G(\boldsymbol{v},\boldsymbol{y}), \ \ \forall \boldsymbol{v} \in \mathbb{R}^{N_1}, \boldsymbol{y} = \mathcal{S}_m(\tilde{\boldsymbol{I}}\boldsymbol{v}). \tag{29}$$

Moreover, a pair $(\boldsymbol{v}^*, \boldsymbol{y}^*)$ is a solution of model (24) if and only if $\boldsymbol{v}^*$ is a solution of model (21) and $\boldsymbol{y}^* = \mathcal{S}_m(\tilde{\boldsymbol{I}}\boldsymbol{v}^*)$.

We remark that the original model (19) presents a three-term single-variable optimization problem (w.r.t. $\boldsymbol{v}$), comprising a convex and differentiable term f, a convex but nondifferentiable term $\iota_{\boldsymbol{q}} \circ \boldsymbol{Q}$, and a nonconvex and nondifferentiable term $\iota_m \circ \tilde{\boldsymbol{I}}$. This model cannot be directly tackled using the PALM algorithm. Resolving such multi-term nondifferentiable optimization models typically necessitates the introduction of proximity operators through subdifferential. When dealing with the summation of two functions φ_1 and φ_2, where at least one function is differentiable, the additivity property of subdifferential (∂) holds:

$$\partial(\varphi_1 + \varphi_2)(\boldsymbol{x}) = \partial\varphi_1(\boldsymbol{x}) + \partial\varphi_2(\boldsymbol{x}).$$

If φ_1 and φ_2 are both nondifferentiable, this property may not hold. The absence of additivity in subdifferential complicates the design of proximity algorithms to solve the original model. Theoretical proof of convergence becomes exceedingly challenging under such circumstances.

The proposed relaxation technique precisely aims to convert the original model into a three-term dual-variable optimization model (w.r.t. $\boldsymbol{v}$ and $\boldsymbol{y}$) in the form of (24), which can be solved via the PALM algorithm. The advantage of such a format lies in its ability to decompose the model into two separate optimization problems, each concerning variables $\boldsymbol{v}$ and $\boldsymbol{y}$ respectively. Moreover, due to the differentiability of H, the subdifferentials corresponding to these two problems exhibit additivity, thus greatly facilitating the theoretical convergence analysis of the PALM algorithm. Furthermore, we demonstrate, through Theorems 3.3 and 3.6, the equivalence of solutions in both formats in the sense of approximation, thereby ensuring the effectiveness of utilizing the PALM algorithm to solve the model.

3.2. Proximal Alternating Linearized Minimization

In this subsection. We employ the Proximal Alternating Linearized Minimization (PALM) algorithm to solve model (24). We verify in the following proposition that the partial gradients $\nabla_{\boldsymbol{v}} H(\boldsymbol{v}, \boldsymbol{y})$ and $\nabla_{\boldsymbol{y}} H(\boldsymbol{v}, \boldsymbol{y})$ of function H defined by (25) are both globally Lipschitz continuous, which is required for the convergence of PALM (Bolte et al., 2014).

Proposition 3.7. *Let function H be defined by (25). Then there exist positive real numbers L_1 and L_2 such that*

$$\begin{aligned}\|\nabla_{\boldsymbol{v}} H(\boldsymbol{v}_1, \boldsymbol{y}) - \nabla_{\boldsymbol{v}} H(\boldsymbol{v}_2, \boldsymbol{y})\|_2 \leqslant L_1 \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2, \\ \forall \boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^{N_1},\ \forall \boldsymbol{y} \in \mathbb{R}^N; \\ \|\nabla_{\boldsymbol{y}} H(\boldsymbol{v}, \boldsymbol{y}_1) - \nabla_{\boldsymbol{y}} H(\boldsymbol{v}, \boldsymbol{y}_2)\|_2 \leqslant L_2 \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2, \\ \forall \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^N,\ \forall \boldsymbol{v} \in \mathbb{R}^{N_1}.\end{aligned}$$

The proof is given in Supplementary A.6. For function $\psi : \mathbb{R}^n \to (-\infty, +\infty]$, we recall that the proximity operator of ψ at $\boldsymbol{x} \in \mathbb{R}^n$ is defined by

$$\operatorname{prox}_{\psi}(\boldsymbol{x}) := \underset{\boldsymbol{u} \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{x}\|_2^2 + \psi(\boldsymbol{u}) \right\}.$$

Then the iterative scheme of PALM to solve model (24) can be given by

$$\begin{cases} \boldsymbol{v}^{k+1} = \operatorname{prox}_{\iota_{\boldsymbol{q}} \circ \boldsymbol{Q}}\big(\boldsymbol{v}^k - \beta_1 \nabla_{\boldsymbol{v}} H(\boldsymbol{v}^k, \boldsymbol{y}^k)\big), \\ \boldsymbol{y}^{k+1} = \mathcal{S}_m\big(\boldsymbol{y}^k - \beta_2 \nabla_{\boldsymbol{y}} H(\boldsymbol{v}^{k+1}, \boldsymbol{y}^k)\big), \end{cases}$$

where β_1 and β_2 are two introduced step-size parameters. As shown in (Bolte et al., 2014), to guarantee the convergence of PALM, it suffices to let $\beta_1 \in \left(0, \frac{1}{L_1}\right)$ and $\beta_2 \in \left(0, \frac{1}{L_2}\right)$, where $L_1 := 2\lambda \|\boldsymbol{h}_2\|_2^2 + \frac{1}{\gamma}$ and $L_2 := \frac{1}{\gamma}$.

To implement the PALM algorithm, we need the closed form of $\operatorname{prox}_{\iota_{\boldsymbol{q}} \circ \boldsymbol{Q}}$, which requires solving the following minimization problem:

$$\underset{\boldsymbol{u} \in \mathbb{R}^{N_1}}{\arg\min} \left\{ \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{v}\|_2^2 + \iota_{\boldsymbol{q}}(\boldsymbol{Q}\boldsymbol{u}) \right\}, \tag{30}$$

for a given $\boldsymbol{v} \in \mathbb{R}^{N_1}$. However, obtaining the closed-form solution of $\operatorname{prox}_{\iota_{\boldsymbol{q}} \circ \boldsymbol{Q}}$ is very tough, due to the complexity of matrix $\boldsymbol{Q}$. Instead, we use the Fixed-Point Proximity Algorithm (FPPA, Micchelli et al. 2011) to solve model (30) iteratively. Before this, we give the closed form of operator $\operatorname{prox}_{\iota_{\boldsymbol{q}}}$. From the definition of proximity operator, it is easy to see that

$$\operatorname{prox}_{\iota_{\boldsymbol{q}}}(\boldsymbol{x}) = \underset{\boldsymbol{u} \in \{\boldsymbol{z} \in \mathbb{R}^{N_2} : \boldsymbol{z} \geqslant \boldsymbol{q}\}}{\arg\min} \left\{ \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{x}\|_2^2 \right\} = \max\{\boldsymbol{x}, \boldsymbol{q}\}.$$

The whole ASMCVaR approach is summarized as Algorithm 1 in Supplementary A.7, with PALM and FPPA in the outer and the inner iterations, respectively.

It is noteworthy that both the proposed relaxation (tailed approximation) technique and the PALM algorithm can be readily extended to various other machine learning problems. To exemplify this, we have introduced two additional modules. Firstly, we apply the proposed relaxation technique and the PALM algorithm to the sparse regression problem, as detailed in Supplementary A.8. Secondly, we have developed a PyTorch module based on this relaxation technique and PALM, facilitating its application to deep learning scenarios. The codes for these two modules are available in the folders *Sparse_Relaxation_Test* and *Pytorch_Demo*, respectively, accessible via the link: `https://github.com/linyizun2024/ASMCVaR`.

4. Experimental Results

We conduct extensive experiments on 6 real-world financial data sets from Kenneth R. French's Data Library[2], whose details are provided in Table 1. They contain monthly price relative sequences of some portfolios formed by different criteria, and one portfolio here is considered as one asset actually. Both FF25 and FF25EU contain 25 portfolios, while the former is built on BE/ME (book equity to market equity) and investment from the US market, the latter is built on ME and prior return from the European market. FF32 consists of 32 portfolios built on BE/ME and investment from the US market, while FF49 consists of 49 industry portfolios from the US market. FF100 and FF100MEOP consist of 100 portfolios from the US market. FF100 is built on ME and BE/ME, while FF100MEOP is built on ME and operating profitability, respectively.

Table 1. Information of 6 real-world monthly benchmark data sets.

Data Set	Region	Time	Months	# Assets
FF25	US	Jul/1971 - May/2023	623	25
FF25EU	EU	Nov/1990 - May/2023	391	25
FF32	US	Jul/1971 - May/2023	623	32
FF49	US	Jul/1971 - May/2023	623	49
FF100	US	Jul/1971 - May/2023	623	100
FF100MEOP	US	Jul/1971 - May/2023	623	100

To assess the effectiveness of the proposed ASMCVaR, we take 9 state-of-the-art PO methods as well as 1 baseline into comparison. They are SSMP (Brodie et al., 2009), SSPO (Lai et al., 2018), SPOLC (Lai et al., 2020), S1, S2, S3 (Luo et al., 2020), Mean-CVaR (Rockafellar & Uryasev, 2000), Mean-CVaR-ℓ_1 (Cheng & Gao, 2015), MT-CVaR (Lai et al., 2022), and the 1/N strategy (baseline, DeMiguel et al. 2009). We also try to run the cutting plane approach (Kobayashi et al., 2021) for Mean-CVaR-ℓ_0 but the program prompts an error that the models are too large in our experiments. There are 3 slightly different versions in (Luo et al., 2020): S1 is deterministic but S2 and S3 are randomized. Thus we average the results of 10 times for S2 and S3. The

[2] `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

parameters for these competitors are set according to their original papers, while those for ASMCVaR will be set in Section 4.1.

We adopt the standard moving-window trading scheme in the literature (Cover, 1991; Li et al., 2016; Lai & Yang, 2023) to assess the investing performance of different methods. Assume we observe $\mathscr{T}$ trading periods for N assets in a financial market. Without loss of generality, we can assume the initial wealth $S^{(0)} = 1$ and the cumulative wealth (CW) $S^{(t)}$ at time t increases (or decreases) with a factor: $S^{(t)} = S^{(t-1)} \cdot ((\mathbf{x}^{(t)})^\top \hat{\boldsymbol{w}}^{(t)})$, where $\mathbf{x}^{(t)} := \boldsymbol{r}_t + \mathbf{1}_N$ denotes the price relative vector at time t and $\hat{\boldsymbol{w}}^{(t)}$ denotes the portfolio for a strategy built on the time window $[t-T : t-1]$. At the beginning where there may be insufficient observations for some strategies, the 1/N strategy can be temporally used. This procedure goes on to the end and forms a backtest CW sequence $\{S^{(t)}\}_{t=0}^{\mathscr{T}}$ that can be used in the assessment. We set for all the methods $T = 60$ as a conventional setting for the time window size in the literature (Goto & Xu, 2015; Ao et al., 2019). All the data sets and codes of this section are accessible via the link: https://github.com/linyizun2024/ASMCVaR/tree/main/Codes_for_Experiments_in_Paper.

4.1. Parameter Setting and Autonomous Sparsity

The model parameters in (21) are set as follows: the confidence level is set as a conventional one $c = 0.99$. The expected return level is empirically set as $\rho = 0.02$ based on observations from real-world financial markets. The approximation parameter is set as $\gamma = 10^{-5}$, which indicates a tight approximation of ℓ_0 constraint. The mixing parameter is set as

$$\lambda = \frac{\sqrt{T}\boldsymbol{h}_1[(N+2):(N+1+T)]^\top(\mathbf{1}_T/T)}{(\boldsymbol{h}_2[1:N]^\top(\mathbf{1}_N/N)-\rho)^2} = \frac{1}{(1-c)\sqrt{T}(\bar{r}-\rho)^2},$$

where $\boldsymbol{h}_1[(N+2):(N+1+T)]$ and $\boldsymbol{h}_2[1:N]$ denote the last T indices of $\boldsymbol{h}_1$ and the first N indices of $\boldsymbol{h}_2$, respectively. $\bar{r}$ is the mean return of the entries of $\boldsymbol{R}$. This setting of λ is a quotient between the CVaR term and the return term in (18) with equal weights $\mathbf{1}_T/T$ and $\mathbf{1}_N/N$. As for the number of selected assets m, we assess 3 different sparsities: $m_1 = 10$, $m_2 = 15$, and $m_3 = 20$.

The algorithm parameters can be conveniently set based on the convergence criteria. For FPPA, we set $\theta = 1.99/\|\tilde{\boldsymbol{Q}}\|_2^2$. Its maximum iteration and relative difference tolerance are set as $MaxIter_1 = 200$ and $tol_1 = 0.001$. For PALM, the learning rates are set as $\beta_1 = \frac{0.99}{L_1}$ and $\beta_2 = \frac{0.99}{L_2}$, respectively. Its maximum iteration and relative difference tolerance are set as $MaxIter_2 = 10^4$ and $tol_2 = 10^{-4}$.

We further investigate the sparsity parameter m and find that the support sets of two different sparsities overlap with a large proportion. To be specific, ASMCVaR produces a portfolio $\hat{\boldsymbol{w}}^{(t,j)}$ for each sparsity m_j and each trade time t. The overlapped proportion can be computed by

$$p_j^{(t)} := \frac{|supp(\hat{\boldsymbol{w}}^{(t,j)}) \cap supp(\hat{\boldsymbol{w}}^{(t,j+1)})|}{|supp(\hat{\boldsymbol{w}}^{(t,j)})|}, \quad j = 1, 2.$$

The sample means $\bar{p}_j$ and the sample standard deviations (STD) $\hat{\sigma}(p_j)$ of $\{p_j^{(t)}\}_{t=1}^{\mathscr{T}}$ on the benchmark data sets are presented in Table 2. More than 89% assets overlap on average between two different sparsities. We call this property autonomous sparsity since a large proportion of assets can remain in the selected asset pool when the portfolio manager adjusts the pool size. It helps to save transaction costs and reduce the managerial burden.

Table 2. Proportion of overlapped assets for ASMCVaR with different sparsities on 6 benchmark data sets.

	FF25	FF25EU	FF32	FF49	FF100	FF100 MEOP
$\bar{p}_1$	0.9115	0.9147	0.8919	0.9468	0.9222	0.9018
$\hat{\sigma}(p_1)$	0.1182	0.1215	0.1307	0.0832	0.1060	0.1154
$\bar{p}_2$	0.9554	0.9553	0.9348	0.9372	0.9543	0.9478
$\hat{\sigma}(p_2)$	0.0808	0.0854	0.0958	0.0825	0.0766	0.0724

4.2. Cumulative Wealth

The CW sequence $\{S^{(t)}\}_{t=0}^{\mathscr{T}}$ represents the investing gain of a method along with time. The final CWs of different methods on 6 benchmark data sets are provided in Table 4. ASMCVaR with any of the 3 sparsity settings outperforms all the other competitors on all 6 data sets. For example, the final CW of ASMCVaR is about 10 times those of Mean-CVaR and Mean-CVaR-ℓ_1 on FF100. Figure 2 in Supplementary A.9 shows the CW plots of different methods on the benchmark data sets. The CW sequences of ASMCVaR ($m = 10$, red color) are steadily over others on most trade times. Besides, ASMCVaR has increasing trends in the long run, indicating that it is growing in wealth during the backtests.

4.3. α Factor

According to the Capital Asset Pricing Model (CAPM, Sharpe 1964), a portfolio cannot exclude the influence from the underlying financial market. To derive a pure excess return to the market, the α factor is proposed (Lintner, 1965) and becomes a popular score for the relative performance to the market. Let r_s and r_m denote the returns for a nontrivial strategy and the Market strategy, respectively. The portfolio return r_s can be decomposed into the market return component r_m and the α factor:

$$\mathbb{E}(r_s) = \beta\mathbb{E}(r_m) + \alpha.$$

Table 3. α factors with p-values of t-tests of different portfolio optimization models on 6 benchmark data sets.

Method	FF25		FF25EU		FF32		FF49		FF100		FF100MEOP	
	α	p-value	α	p-value	α	p-value	α	p-value	α	p-value	α	p-value
1/N	0	0.4318	-0.0031	1	0.0001	0.3782	0	0.5487	-0.0004	0.9427	-0.0002	0.9083
SSMP	0.0002	0.4465	-0.0027	0.9992	-0.0028	0.9744	0.0017	0.1935	-0.0068	0.9999	-0.0051	0.9983
SSPO	-0.0024	0.9440	-0.0098	1	-0.0046	0.9994	-0.0056	0.9278	-0.0100	0.9999	-0.0071	0.9996
SPOLC	-0.0032	0.9984	-0.0103	1	-0.0011	0.8050	-0.0037	0.9008	-0.0073	1	-0.0050	0.9977
S1	-0.0027	0.9701	-0.0104	1	-0.0049	0.9997	-0.0062	0.9463	-0.0110	1	-0.0069	0.9991
S2	-0.0029	0.9740	-0.0101	1	-0.0050	0.9996	-0.0062	0.9475	-0.0106	1	-0.0072	0.9994
S3	-0.0028	0.9738	-0.0103	1	-0.0049	0.9997	-0.0062	0.9465	-0.0110	1	-0.0068	0.9991
Mean-CVaR	0.0009	0.1133	-0.0014	0.9216	0.0019	0.0183	0.0025	0.0098	-0.0006	0.7222	0.0017	0.0405
Mean-CVaR-ℓ_1	0.0015	0.0339	-0.0011	0.8947	0.0017	0.0300	0.0027	0.0065	0	0.4987	0.0006	0.2969
MT-CVaR	-0.0041	0.9984	-0.0090	1	-0.0067	1	-0.0044	0.8979	-0.0105	1	-0.0050	0.9891
ASMCVaR($m = 10$)	**0.0022**	0.0006	**0.0027**	0	**0.0027**	0.0002	**0.0029**	0.0049	**0.0023**	0.0038	**0.0020**	0.0060
ASMCVaR($m = 15$)	**0.0019**	0.0029	**0.0027**	0	**0.0025**	0.0001	0.0024	0.0112	**0.0025**	0.0012	**0.0017**	0.0109
ASMCVaR($m = 20$)	**0.0019**	0.0017	**0.0025**	0	**0.0024**	0.0001	0.0023	0.0098	**0.0024**	0.0010	0.0016	0.0114

Table 4. Final cumulative wealths of different portfolio optimization models on 6 benchmark data sets.

Method	FF25	FF25EU	FF32	FF49	FF100	FF100 MEOP
1/N	355.98	13.05	424.42	235.48	364.87	348.70
SSMP	248.67	13.47	158.98	186.79	10.09	26.06
SSPO	129.35	1.22	30.20	1.33	0.89	3.90
SPOLC	57.53	0.96	169.60	5.44	2.39	8.23
S1	100.76	1.08	29.47	1.09	0.54	4.31
S2	83.10	1.19	25.93	1.15	0.66	3.55
S3	96.26	1.11	29.45	1.09	0.55	4.38
Mean-CVaR	333.41	18.60	331.14	150.15	70.31	306.40
Mean-CVaR-ℓ_1	440.32	18.58	332.55	178.93	102.11	163.77
MT-CVaR	34.58	1.53	7.96	4.93	0.61	11.66
ASMCVaR($m = 10$)	**973.23**	**122.82**	**1122.76**	**758.66**	**966.32**	**783.47**
ASMCVaR($m = 15$)	**827.18**	**123.13**	**1303.80**	**694.83**	**1126**	**712.45**
ASMCVaR($m = 20$)	**872.03**	**113.28**	**1317.01**	**767.61**	**1117.20**	**714.49**

β and α can be computed by the linear regression:

$$\hat{\beta} = \frac{\hat{c}\left(r_s, r_m\right)}{\hat{\sigma}^2\left(r_m\right)}, \quad \hat{\alpha} = \bar{r}_s - \hat{\beta}\bar{r}_m,$$

where $\hat{c}(\cdot,\cdot)$ and $\hat{\sigma}^2(\cdot)$ are the sample covariance and variance, respectively. $\bar{r}_s$ denotes the sample mean, which can be computed by $\bar{r}_s = \frac{1}{\mathscr{T}}\sum_{t=1}^{\mathscr{T}} r_s^{(t)}$, where $r_s^{(t)} = (\mathbf{x}^{(t)})^\top \hat{\boldsymbol{w}}_s^{(t)} - 1$ and $\hat{\boldsymbol{w}}_s^{(t)}$ is the portfolio of this nontrivial strategy at time t. $\bar{r}_m$ can be computed likewise. We adopt the uniform-buy-and-hold strategy (Li et al., 2016) as the conventional representative of the market return. To test whether $\alpha > 0$ is statistically significant, the right-tailed t-test can also be implemented.

Table 3 shows the α factors with p-values of t-tests of different methods. ASMCVaR($m = 10$) achieves the highest αs among all the compared methods on all 6 data sets. Moreover, ASMCVaR($m = 10$) achieves positive αs at the significance level of 1% in all the cases, since all the p-values are smaller than 1%. Hence ASMCVaR obtains positive excess returns to the market with high statistical significance.

4.4. Sharpe Ratio

Since an investor usually has to take a higher risk to obtain a higher return in financial markets, it is necessary to develop risk-adjusted returns to evaluate this balancing performance. The Sharpe ratio (SR, Sharpe 1966) is such a metric that takes the ratio between return and risk of a strategy:

$$SR = \frac{\bar{r}_s - r_f}{\hat{\sigma}\left(r_s\right)},$$

where the risk-free return is set as $r_f = 0$ in this paper. The (monthly) SRs for different methods on the 6 data sets are shown in Table 5. ASMCVaR with $m = 10$, $m = 15$ or $m = 20$ achieves the highest SRs among all the competitors on all the data sets. For example, the SRs of ASMCVaR are about 45.7% higher than those of Mean-CVaR and Mean-CVaR-ℓ_1 on FF25EU. These results indicate that ASMCVaR performs well in balancing return and risk.

Table 5. Sharpe ratios of different portfolio optimization models on 6 benchmark data sets.

Method	FF25	FF25EU	FF32	FF49	FF100	FF100 MEOP
1/N	0.2278	0.1576	0.2236	0.2059	0.2089	0.2077
SSMP	0.1935	0.1598	0.1536	0.1659	0.0883	0.1085
SSPO	0.1545	0.0412	0.1182	0.0588	0.0425	0.0685
SPOLC	0.1453	0.0316	0.1735	0.0753	0.0563	0.0864
S1	0.1498	0.0370	0.1169	0.0559	0.0327	0.0709
S2	0.1446	0.0408	0.1137	0.0571	0.0367	0.0670
S3	0.1486	0.0381	0.1170	0.0559	0.0331	0.0712
Mean-CVaR	0.2305	0.1803	0.2375	0.2220	0.1738	0.2204
Mean-CVaR-ℓ_1	0.2392	0.1862	0.2330	0.2262	0.1815	0.1939
MT-CVaR	0.1251	0.0505	0.0844	0.0767	0.0325	0.0911
ASMCVaR($m = 10$)	**0.2610**	**0.2758**	**0.2638**	**0.2339**	**0.2446**	**0.2363**
ASMCVaR($m = 15$)	**0.2541**	**0.2765**	**0.2615**	**0.2286**	**0.2503**	**0.2326**
ASMCVaR($m = 20$)	**0.2558**	**0.2713**	**0.2603**	**0.2300**	**0.2501**	**0.2322**

4.5. Transaction Cost

The transaction cost with portfolio change is inevitable in practical portfolio management. We adopt the proportional transaction cost model (Blum & Kalai, 1999; Lai & Yang, 2023) to compute the CW at the beginning of the t-th trade

time:

$$S^{(\mathscr{T})}=S^{(0)}\prod_{t=1}^{\mathscr{T}}\left[\left((\mathbf{x}^{(t)})^\top\hat{\boldsymbol{w}}^{(t)}\right)\left(1-\frac{\nu}{2}\sum_{i=1}^{N}\left|\hat{w}_i^{(t)}-\tilde{w}_i^{(t-1)}\right|\right)\right],$$

$$\tilde{w}_i^{(t-1)}=\frac{\hat{w}_i^{(t-1)}\cdot \mathrm{x}_i^{(t-1)}}{(\mathbf{x}^{(t-1)})^\top\hat{\boldsymbol{w}}^{(t-1)}},$$

where $\tilde{w}_i^{(t-1)}$ denotes the evolved portfolio weight of the i-th asset at the end of the $(t-1)$-th trade time, and $\tilde{\boldsymbol{w}}^{(0)}$ can be initialized as $\mathbf{0}_N$. ν is the bidirectional transaction cost rate, which is examined in the range $[0, 0.5\%]$. When the cost rates of buying and selling are the same, changing the evolved portfolio $\tilde{\boldsymbol{w}}^{(t-1)}$ to the next portfolio $\tilde{\boldsymbol{w}}^{(t)}$ yields a proportional transaction cost of $\frac{\nu}{2}\sum_{i=1}^{N}\left|\hat{w}_i^{(t)}-\tilde{w}_i^{(t-1)}\right|$.

When ν changes from 0 to 0.5%, the final CWs for different methods are plotted in Figure 3 of Supplementary A.10. ASMCVaR outperforms all the other competitors on all the data sets with all ν, showing a good performance against the transaction cost. Note that the 1/N strategy is well known for its little turnover rate mechanism, but ASMCVaR still has an advantage over 1/N, especially with low and medium cost rates. Moreover, if a mutual fund has sufficient capital, then the transaction cost rate ν can be negotiated to be low. Hence ASMCVaR is effective in practical portfolio management with transaction cost.

5. Conclusion

We develop an autonomous sparse mean-CVaR (ASMCVaR) model for portfolio optimization. It aims to optimize a mean-CVaR objective function under the ℓ_0 and the simplex constraints for exact and feasible asset selection. It is an NP-hard problem and the existing combinatorial approach requires a high computational cost. From a markedly different perspective, we propose to convert the ℓ_0 constraint into an indicator function and handle it with a tailed approximation. By this way, ASMCVaR can approximate the ℓ_0-constrained mean-CVaR model to arbitrary accuracy. We also develop a proximal alternating linearized minimization algorithm, coupled with a nested fixed-point proximity algorithm (both convergent) to iteratively solve ASMCVaR. It has a good property of autonomous sparsity because a large proportion of assets can remain in the selected asset pool when adjusting the pool size. Experimental results show that ASMCVaR achieves state-of-the-art performance in several common evaluating metrics of investing performance and risk control.

A direct impact of ASMCVaR is to provide a novel tailed approximation approach to the long-standing NP-hard ℓ_0-constrained problem. It can be conveniently transferred to other related ℓ_0-constrained problems via similar establishment. Further improvements on ASMCVaR may lie in accelerating the solving algorithm with momentum methods, generalizing the objective function to a broader class of functions, or considering more constraints with practical sense.

Acknowledgements

The authors thank the anonymous reviewers for their constructive comments and valuable suggestions in improving this paper. This work was supported in part by National Natural Science Foundation of China under Grant 62176103, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110541, and in part by the Science and Technology Planning Project of Guangzhou under Grants 2024A04J3940 and 2024A04J9896.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are no potential societal consequences of our work that we feel must be specifically highlighted here.

References

Ao, M., Li, Y., and Zheng, X. Approaching mean-variance efficiency for large portfolios. *The Review of Financial Studies*, 32(7):2890–2919, 2019.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

Ban, G.-Y., Karoui, N. E., and Lim, A. E. B. Machine learning and portfolio optimization. *Management Science*, 64(3):1136–1154, Mar. 2018.

Bertsimas, D. and Cory-Wright, R. A scalable algorithm for sparse portfolio selection. *INFORMS Journal on Computing*, 34(3):1489–1511, 2022.

Blum, A. and Kalai, A. Universal portfolios with and without transaction costs. *Machine Learning*, 35(3):193–205, 1999.

Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

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

Chao, R. O., Kavadias, S., and Gaimon, C. Revenue driven resource allocation: Funding authority, incentives, and new product development portfolio management. *Management Science*, 55(9):1556–1569, 2009.

Cheng, R. and Gao, J. On cardinality constrained mean-cvar portfolio optimization. In *The 27th Chinese Control and Decision Conference (2015 CCDC)*, pp. 1074–1079, 2015.

Cover, T. M. Universal portfolios. *Mathematical Finance*, 1 (1):1–29, Jan. 1991.

DeMiguel, V., Garlappi, L., and Uppal, R. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, May 2009.

Dimmock, S. G., Wang, N., and Yang, J. The endowment model and modern portfolio theory. *Management Science*, pp. 1–26, Apr. 2023. doi: 10.1287/mnsc.2023.4759.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Annals of Statistics*, 32(2):407–451, 2004.

Goto, S. and Xu, Y. Improving mean variance optimization through sparse hedging restrictions. *The Journal of Financial and Quantitative Analysis*, 50(6):1415–1441, 2015.

Gotoh, J.-Y. and Takeda, A. On the role of norm constraints in portfolio selection. *Computational Management Science*, 8:323–353, 2011.

Günlük, O. and Linderoth, J. Perspective reformulations of mixed integer nonlinear programs with indicator variables. *Mathematical Programming*, 124(1):183–205, 2010.

Jorion, P. *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, New York, 1997.

Kobayashi, K., Takano, Y., and Nakata, K. Bilevel cutting-plane algorithm for cardinality-constrained mean-cvar portfolio optimization. *Journal of Global Optimization*, 81(2):493–528, 2021.

Lai, Z.-R. and Yang, H. A survey on gaps between mean-variance approach and exponential growth rate approach for portfolio optimization. *ACM Computing Surveys*, 55 (2):1–36, Mar. 2023. Article No. 25.

Lai, Z.-R., Yang, P.-Y., Fang, L., and Wu, X. Short-term sparse portfolio optimization based on alternating direction method of multipliers. *Journal of Machine Learning Research*, 19(63):1–28, Oct. 2018.

Lai, Z.-R., Tan, L., Wu, X., and Fang, L. Loss control with rank-one covariance estimate for short-term portfolio optimization. *Journal of Machine Learning Research*, 21 (97):1–37, Jun. 2020.

Lai, Z.-R., Li, C., Wu, X., Guan, Q., and Fang, L. Multitrend conditional value at risk for portfolio optimization. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022. doi: 10.1109/TNNLS.2022.3183891.

Li, B., Sahoo, D., and Hoi, S. C. OLPS: a toolbox for online portfolio selection. *Journal of Machine Learning Research*, 17(1):1242–1246, 2016.

Lintner, J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1):13–37, Feb. 1965.

Liu, H. and Loewenstein, M. Market crashes, correlated illiquidity, and portfolio choice. *Management Science*, 59 (3):715–732, 2013.

Luo, Z., Yu, X., Xiu, N., and Wang, X. Closed-form solutions for short-term sparse portfolio optimization. *Optimization*, 2020.

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

Micchelli, C. A., Shen, L., and Xu, Y. Proximity algorithms for image models: denoising. *Inverse Problems*, 27(4): 045009, 2011.

Ogryczak, W. and Ruszczyński, A. Dual stochastic dominance and related mean-risk models. *SIAM Journal on Optimization*, 13:60–78, 2002.

Rockafellar, R. T. and Uryasev, S. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Sharpe, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, Sep. 1964.

Sharpe, W. F. Mutual fund performance. *Journal of Business*, 39(1):119–138, Jan. 1966.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1): 267–288, 1996.

A. Supplementary Materials

A.1. Proof of Lemma 3.2

Proof. For the case $\|\boldsymbol{w}^*\|_0 \leqslant m$, it is obvious that (22) holds, since $w_j^* = 0$ for all $j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}^*}^m$. We then consider the case $\|\boldsymbol{w}^*\|_0 > m$. In this case, we construct an approximation $\tilde{\boldsymbol{v}} := (\tilde{\boldsymbol{w}}^\top, \tilde{\tau}, \tilde{\boldsymbol{z}}^\top)^\top \in \mathbb{R}^{N_1}$ of $\boldsymbol{v}^* := ((\boldsymbol{w}^*)^\top, \tau^*, (\boldsymbol{z}^*)^\top)^\top$ such that it satisfies the constraints of (17), that is, the constraints of (14). Note that $\boldsymbol{v}^*$ satisfies that $\boldsymbol{Q}\boldsymbol{v}^* \geqslant \boldsymbol{q}$. We let $\tilde{w}_j = w_j^*$ for $j \in J_{\boldsymbol{w}^*}^m \backslash \{j_1\}$, $\tilde{w}_j = 0$ for $j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}^*}^m$ and $\tilde{w}_{j_1} = w_{j_1}^* + \sum_{j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}^*}^m} w_j^*$. Then $\tilde{\boldsymbol{w}} \in \Delta$ and $\|\tilde{\boldsymbol{w}}\|_0 \leqslant m$. For $\tilde{\boldsymbol{z}}$ and $\tilde{\tau}$, we define $\delta := \max_{j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}^*}^m} \{w_j^*\}$ and $\zeta := \max_{i \in \mathbb{N}_T, j \in \mathbb{N}_N} \{|r_{i,j} - r_{i,j_1}|\}$, where $r_{i,j}$ is the (i,j)th component of matrix $\boldsymbol{R}$ in model (21), and then let $\tilde{\boldsymbol{z}} = \boldsymbol{z}^*$, $\tilde{\tau} = \tau^* + \delta\zeta$. It is obvious that $\tilde{\boldsymbol{z}} \geqslant \boldsymbol{0}_T$ since $\boldsymbol{z}^* \geqslant \boldsymbol{0}_T$. To show that $\boldsymbol{Q}\tilde{\boldsymbol{v}} \geqslant \boldsymbol{q}$, it remains to be shown that

$$\tilde{\tau}\boldsymbol{1}_T \geqslant -\tilde{\boldsymbol{z}} - \boldsymbol{R}\tilde{\boldsymbol{w}}. \tag{31}$$

Note that $\boldsymbol{R}\boldsymbol{w}^* - \boldsymbol{R}\tilde{\boldsymbol{w}} = \sum_{j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}^*}^m} w_j^* \left(\boldsymbol{r}^{(j)} - \boldsymbol{r}^{(j_1)}\right)$, where $\boldsymbol{r}^{(j)}$ is the jth column of $\boldsymbol{R}$. From the definitions of δ and ζ, we have that

$$\tilde{\tau}\boldsymbol{1}_T \geqslant \tau^*\boldsymbol{1}_T + \sum_{j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}^*}^m} w_j^* \left(\boldsymbol{r}^{(j)} - \boldsymbol{r}^{(j_1)}\right) = \tau^*\boldsymbol{1}_T + \boldsymbol{R}\boldsymbol{w}^* - \boldsymbol{R}\tilde{\boldsymbol{w}}. \tag{32}$$

Then (31) follows from (32) and the fact $\tau^*\boldsymbol{1}_T \geqslant -\boldsymbol{z}^* - \boldsymbol{R}\boldsymbol{w}^*$.

Now we know that $\tilde{\iota}_{m,\gamma}(\tilde{\boldsymbol{w}}) = 0$ and $\iota_{\boldsymbol{q}}(\boldsymbol{Q}\tilde{\boldsymbol{v}}) = 0$. Hence

$$\Psi(\tilde{\boldsymbol{v}}) = f(\tilde{\boldsymbol{v}}). \tag{33}$$

The fact $\boldsymbol{w}^* \in \Delta$ gives that $\|\boldsymbol{w}^*\|_2 \leqslant \sqrt{\|\boldsymbol{w}^*\|_1} = 1$, where $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the ℓ_2 and ℓ_1 norms, respectively. Similarly, we also have $\|\tilde{\boldsymbol{w}}\|_2 \leqslant 1$. In addition,

$$\|\boldsymbol{w}^* - \tilde{\boldsymbol{w}}\|_2^2 = \left(\sum_{j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}^*}^m} w_j^*\right)^2 + \sum_{j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}^*}^m} (w_j^*)^2 \leqslant [(N-m)^2 + (N-m)]\delta^2$$

Then it follows from the definition of function f in (18) that

$$\begin{aligned} |f(\boldsymbol{v}^*) - f(\tilde{\boldsymbol{v}})| &= |(\tau^* - \tilde{\tau}) + \lambda(\boldsymbol{\mu}^\top \boldsymbol{w}^* - \rho)^2 - \lambda(\boldsymbol{\mu}^\top \tilde{\boldsymbol{w}} - \rho)^2| \\ &\leqslant \delta\zeta + \lambda|\boldsymbol{\mu}^\top(\boldsymbol{w}^* + \tilde{\boldsymbol{w}}) - 2\rho)| \cdot |\boldsymbol{\mu}^\top(\boldsymbol{w}^* - \tilde{\boldsymbol{w}})| \\ &\leqslant \delta\zeta + \lambda\left(2\|\boldsymbol{\mu}\|_2 + 2\rho\right)\|\boldsymbol{\mu}\|_2\|\boldsymbol{w}^* - \tilde{\boldsymbol{w}}\|_2 \\ &\leqslant \tilde{L}\delta \leqslant \tilde{L}, \end{aligned} \tag{34}$$

where

$$\tilde{L} := \zeta + 2\lambda\sqrt{[(N-m)^2 + (N-m)]}\left(\|\boldsymbol{\mu}\|_2^2 + \rho\|\boldsymbol{\mu}\|_2\right). \tag{35}$$

This together with (33) and the fact $\Psi(\boldsymbol{v}^*) \leqslant \Psi(\tilde{\boldsymbol{v}})$ yields that

$$f(\boldsymbol{v}^*) + \tilde{L} \geqslant f(\tilde{\boldsymbol{v}}) \geqslant \Psi(\boldsymbol{v}^*) = f(\boldsymbol{v}^*) + \tilde{\iota}_{m,\gamma}(\boldsymbol{w}^*) \geqslant f(\boldsymbol{v}^*) + \frac{1}{2\gamma}\delta^2.$$

Thus $\delta \leqslant \sqrt{2\tilde{L}\gamma}$, which implies the desired result. □

A.2. Proof of Theorem 3.3

Proof. For the case $\|\tilde{\boldsymbol{I}}\boldsymbol{v}^*\|_0 \leqslant m$, since $\tilde{\iota}_{m,\gamma}(\tilde{\boldsymbol{I}}\boldsymbol{v}^*) = \iota_m(\tilde{\boldsymbol{I}}\boldsymbol{v}^{**}) = 0$ and $\iota_{\boldsymbol{q}}(\boldsymbol{Q}\boldsymbol{v}^*) = \iota_{\boldsymbol{q}}(\boldsymbol{Q}\boldsymbol{v}^{**}) = 0$, it is obvious that $f(\boldsymbol{v}^*) = f(\boldsymbol{v}^{**})$, and hence (23) holds. For the case $\|\tilde{\boldsymbol{I}}\boldsymbol{v}^*\|_0 > m$, we define $\tilde{\boldsymbol{v}}$ as an approximation of $\boldsymbol{v}^*$, following the same definition as in the proof of Lemma 3.2. Then $\boldsymbol{Q}\tilde{\boldsymbol{v}} \geqslant \boldsymbol{q}$ and $\|\tilde{\boldsymbol{I}}\tilde{\boldsymbol{v}}\|_0 \leqslant m$. Further, $f(\boldsymbol{v}^{**}) \leqslant f(\tilde{\boldsymbol{v}})$. By Lemma 3.2 and the inequality (34) in its proof, we know that $|f(\boldsymbol{v}^*) - f(\tilde{\boldsymbol{v}})| \leqslant \tilde{L}\delta \leqslant \sqrt{2\tilde{L}^3\gamma}$, where $\tilde{L}$ is defined by (35) and $\delta := \max_{j \in \mathbb{N}_N \backslash J_{\boldsymbol{w}^*}^m} \{w_j^*\}$. Thus

$$f(\boldsymbol{v}^{**}) \leqslant f(\tilde{\boldsymbol{v}}) \leqslant f(\boldsymbol{v}^*) + \sqrt{2\tilde{L}^3\gamma}. \tag{36}$$

The prerequisite $\boldsymbol{v}^*$ and $\boldsymbol{v}^{**}$ are solutions of models (21) and (19) gives that $\Psi(\boldsymbol{v}^*) \leqslant \Psi(\boldsymbol{v}^{**})$ and $\Psi(\boldsymbol{v}^{**}) = f(\boldsymbol{v}^{**})$. Hence

$$f(\boldsymbol{v}^*) \leqslant \Psi(\boldsymbol{v}^*) \leqslant f(\boldsymbol{v}^{**}). \tag{37}$$

Now (23) can be obtained by (36) and (37) directly. This completes the proof. □

A.3. Proof of Lemma 3.4

Proof. Since $(\boldsymbol{v}^*, \boldsymbol{y}^*)$ is a solution of model (24), we have that $\|\boldsymbol{y}^*\|_0 \leqslant m$ and

$$G(\boldsymbol{v}^*, \boldsymbol{y}^*) \leqslant G(\boldsymbol{v}^*, \boldsymbol{y}),\ \forall \boldsymbol{y} \in \mathbb{R}^N. \tag{38}$$

Subtracting $f(\boldsymbol{v}^*) + \iota_{\boldsymbol{q}}(\boldsymbol{Q}\boldsymbol{v}^*)$ on both sides of (38) yields that

$$\frac{1}{2\gamma}\|\widetilde{\boldsymbol{I}}\boldsymbol{v}^* - \boldsymbol{y}^*\|_2^2 + \iota_m(\boldsymbol{y}^*) \leqslant \frac{1}{2\gamma}\|\widetilde{\boldsymbol{I}}\boldsymbol{v}^* - \boldsymbol{y}\|_2^2 + \iota_m(\boldsymbol{y}),\ \forall \boldsymbol{y} \in \mathbb{R}^N. \tag{39}$$

Let $\boldsymbol{w}^* := \tilde{\boldsymbol{I}}\boldsymbol{v}^*$. Then (39) implies that

$$\|\boldsymbol{w}^* - \boldsymbol{y}^*\|_2^2 \leqslant \|\boldsymbol{w}^* - \boldsymbol{y}\|_2^2,\ \forall \boldsymbol{y} \in \Omega_m. \tag{40}$$

We consider two cases for $\boldsymbol{w}^*$. If $\boldsymbol{w}^* \in \Omega_m$, (40) implies that $\boldsymbol{y}^* = \boldsymbol{w}^*$, that is, $\boldsymbol{y}^* = \mathcal{S}_m(\boldsymbol{w}^*)$.

We then consider the case $\boldsymbol{w}^* \notin \Omega_m$. In this case, we have $\|\boldsymbol{y}^*\|_0 = m$. Suppose not, then $\|\boldsymbol{y}^*\|_0 < m$. Since $\boldsymbol{w}^* \notin \Omega_m$, we know that $\boldsymbol{w}^*$ has at least $m+1$ nonzero components. Hence there exists some $j' \in \mathbb{N}_N$ such that $y_{j'}^* = 0$ and $w_{j'}^* \neq 0$. By letting $\boldsymbol{y} \in \Omega_m$ such that

$$y_{j'} = w_{j'}^* \ \text{ and } \ y_j = y_j^* \text{ for } j \in \mathbb{N}_N \backslash \{j'\}, \tag{41}$$

then

$$\|\boldsymbol{w}^* - \boldsymbol{y}^*\|^2 - \|\boldsymbol{w}^* - \boldsymbol{y}\|^2 = (w_{j'}^* - y_{j'}^*)^2 > 0, \tag{42}$$

which contradicts (40). Thus $\|\boldsymbol{y}^*\|_0 = m$. Note that the components in $\boldsymbol{w}^*$ may not be distinct. We let $J_1, J_2, J_3 \subset \mathbb{N}_N$ be three index sets such that $|w_j^*| > |w_{j_m}^*|$ for all $j \in J_1$, $|w_j^*| = |w_{j_m}^*|$ for all $j \in J_2$, and $|w_j^*| < |w_{j_m}^*|$ for all $j \in J_3$, respectively.

(i) We first prove that $y_j^* = w_j^*$ for all $j \in J_1$ if $J_1 \neq \varnothing$. To this end, we show that $y_j^* \neq 0$ for all $j \in J_1$. Assume that there exists some $j' \in J_1$ such that $y_{j'}^* = 0$. Since $\|\boldsymbol{y}^*\|_0 = m$, there must be some $j'' \in \mathbb{N}_N \backslash J_1$ such that $y_{j''}^* \neq 0$. Note that $|w_{j'}^*| > |w_{j''}^*|$. By letting $\boldsymbol{y} \in \Omega_m$ such that

$$y_{j'} = w_{j'}^*,\ y_{j''} = 0 \text{ and } y_j = y_j^* \text{ for } j \in \mathbb{N}_N \backslash \{j', j''\}, \tag{43}$$

then

$$\|\boldsymbol{w}^* - \boldsymbol{y}^*\|_2^2 - \|\boldsymbol{w}^* - \boldsymbol{y}\|_2^2 = (w_{j'}^*)^2 + (w_{j''}^* - y_{j''}^*)^2 - (w_{j''}^*)^2 > 0, \tag{44}$$

which contradicts (40). Hence $y_j^* \neq 0$ for all $j \in J_1$. Now we are able to prove by contradiction that $y_j^* = w_j^*$ for all $j \in J_1$. Assume that there exists $j' \in J_1$ such that $y_j^* \neq w_{j'}^*$. Since $y_{j'}^* \neq 0$, letting $\boldsymbol{y} \in \mathbb{R}^N$ be given by (41), then $\boldsymbol{y} \in \Omega_m$ and (42) holds, which contradicts (40). Hence $y_j^* = w_j^*$ for all $j \in J_1$.

(ii) We next prove that $y_j^* = 0$ for all $j \in J_3$ if $J_3 \neq \varnothing$. Otherwise, there exists $j'' \in J_3$ such that $y_{j''}^* \neq 0$. Note that the number of elements in $J_1 \cup J_2 = \mathbb{N}_N \backslash J_3$ is greater than or equal to m. To guarantee that $\|\boldsymbol{y}^*\|_0 = m$, there must be some $j' \in \mathbb{N}_N \backslash J_3$ such that $y_{j'}^* = 0$. In this case, we let $\boldsymbol{y} \in \Omega_m$ be given by (43). Note that $|w_{j'}^*| > |w_{j''}^*|$. We have that (44) holds, which contradicts (40). Hence $y_j^* = 0$ for all $j \in J_3$.

Note that $J_1 \subsetneqq J_{\boldsymbol{w}^*}^m$, $J_{\boldsymbol{w}^*}^m \backslash J_1 \subset J_2$. In addition, $\boldsymbol{w}^* \notin \Omega_m$ implies that $w_j^* \neq 0$ for all $j \in J_1 \cup J_2$. Let $m_1 \in \{0, 1, \ldots, m-1\}$ be the number of elements in J_1. The fact $\|\boldsymbol{y}^*\|_0 = m$ together with (i) and (ii) implies that there must be an index set $J_2' \subset J_2$ with $(m - m_1)$ components such that $y_j^* \neq 0$ for $j \in J_2'$ and $y_j^* = 0$ for $j \in J_2 \backslash J_2'$. As shown in (i), we can also prove that $y_j^* = w_j^*$ for $j \in J_2'$. Now by setting $J = J_1 \cup J_2'$, then it is an m-LAV index set of $\boldsymbol{w}$. In addition, we see from (27) that $\boldsymbol{y}^* = \mathcal{S}_m(\widetilde{\boldsymbol{I}}\boldsymbol{v}^*)$, which completes the proof. □

A.4. Proof of Lemma 3.5

Proof. Let $\boldsymbol{w} := \widetilde{\boldsymbol{I}}\boldsymbol{v}$. We first prove that

$$\|\tilde{\boldsymbol{y}} - \boldsymbol{w}\|_2^2 \leqslant \|\boldsymbol{y} - \boldsymbol{w}\|_2^2, \ \forall \boldsymbol{y} \in \Omega_m. \tag{45}$$

If $\boldsymbol{w} \in \Omega_m$, then it is easy to see from the definition (27) of operator $\mathcal{S}_m$ that $\tilde{\boldsymbol{y}} = \boldsymbol{w}$ and hence (45) holds. For the case $\boldsymbol{w} \notin \Omega_m$, since $\tilde{\boldsymbol{y}} = \mathcal{S}_m(\widetilde{\boldsymbol{I}}\boldsymbol{v})$, we know that $\tilde{\boldsymbol{y}} \in \Omega_m$ and

$$\|\tilde{\boldsymbol{y}} - \boldsymbol{w}\|_2^2 = \sum_{j \in \mathbb{N}_N \setminus J_{\boldsymbol{w}}^m} w_j^2. \tag{46}$$

For any $\boldsymbol{y} \in \Omega_m$, there exists an index set $J_{\boldsymbol{y}} \subset \mathbb{N}_N$ with m elements such that $y_j = 0$ for all $j \in \mathbb{N}_N \setminus J_{\boldsymbol{y}}$. Since $J_{\boldsymbol{w}}^m$ contains the m largest components of $\boldsymbol{w}$ in absolute value, we have that $\sum_{j \in J_{\boldsymbol{w}}^m} w_j^2 \geqslant \sum_{j \in J_{\boldsymbol{y}}} w_j^2$, which together with (46) implies that

$$\|\boldsymbol{y} - \boldsymbol{w}\|_2^2 \geqslant \sum_{j \in \mathbb{N}_N \setminus J_{\boldsymbol{y}}} w_j^2 \geqslant \sum_{j \in \mathbb{N}_N \setminus J_{\boldsymbol{w}}^m} w_j^2 = \|\tilde{\boldsymbol{y}} - \boldsymbol{w}\|_2^2.$$

This completes the proof of (45). Now by (45), we have

$$\frac{1}{2\gamma}\|\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{I}}\boldsymbol{v}\|_2^2 + \iota_m(\tilde{\boldsymbol{y}}) \leqslant \frac{1}{2\gamma}\|\boldsymbol{y} - \widetilde{\boldsymbol{I}}\boldsymbol{v}\|_2^2 + \iota_m(\boldsymbol{y}), \ \forall \boldsymbol{y} \in \mathbb{R}^N. \tag{47}$$

Adding $f(\boldsymbol{v}) + \iota_{\boldsymbol{q}}(\boldsymbol{Q}\boldsymbol{v})$ on both sides of (47) yields (28) immediately. □

A.5. Proof of Theorem 3.6

Proof. Let $\boldsymbol{w} := \tilde{\boldsymbol{I}}\boldsymbol{v}$. The definition of $\tilde{\iota}_{m,\gamma}$ in (20) implies that $\tilde{\iota}_{m,\gamma}(\boldsymbol{w}) = \frac{1}{2\gamma}\|\boldsymbol{w} - \mathcal{S}_m(\boldsymbol{w})\|_2^2$. Then for $\boldsymbol{y} = \mathcal{S}_m(\boldsymbol{w})$, we know that $\iota_m(\boldsymbol{y}) = 0$ and $\tilde{\iota}_{m,\gamma}(\boldsymbol{w}) = \frac{1}{2\gamma}\|\boldsymbol{w} - \boldsymbol{y}\|_2^2$. Hence $\Psi(\boldsymbol{v}) = G(\boldsymbol{v}, \boldsymbol{y})$.

Suppose that $(\boldsymbol{v}^*, \boldsymbol{y}^*)$ is a solution of model (24). From Lemma 3.4, we know that $\boldsymbol{y}^* = \mathcal{S}_m(\widetilde{\boldsymbol{I}}\boldsymbol{v}^*)$, and hence $\Psi(\boldsymbol{v}^*) = G(\boldsymbol{v}^*, \boldsymbol{y}^*)$. If $\boldsymbol{v}^*$ is not a solution of model (21), then there exists $\tilde{\boldsymbol{v}} \in \mathbb{R}^{N_1}$ such that $\Psi(\tilde{\boldsymbol{v}}) < \Psi(\boldsymbol{v}^*)$. In this case, we have that $G(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{y}}) = \Psi(\tilde{\boldsymbol{v}}) < \Psi(\boldsymbol{v}^*) = G(\boldsymbol{v}^*, \boldsymbol{y}^*)$, for any $\tilde{\boldsymbol{y}} = \mathcal{S}_m(\widetilde{\boldsymbol{I}}\tilde{\boldsymbol{v}})$, which contradicts the assumption that $(\boldsymbol{v}^*, \boldsymbol{y}^*)$ is a solution of model (24). This proves that $\boldsymbol{v}^*$ is a solution of model (21).

Conversely, suppose that $\boldsymbol{v}^*$ is a solution of model (21) and $\boldsymbol{y}^* = \mathcal{S}_m(\widetilde{\boldsymbol{I}}\boldsymbol{v}^*)$. We confirm that $(\boldsymbol{v}^*, \boldsymbol{y}^*)$ is a solution of model (24). If it is not true, according to Lemma 3.5, we can find vectors $\bar{\boldsymbol{v}} \in \mathbb{R}^{N_1}$ and $\bar{\boldsymbol{y}} = \mathcal{S}_m(\widetilde{\boldsymbol{I}}\bar{\boldsymbol{v}})$ such that $G(\bar{\boldsymbol{v}}, \bar{\boldsymbol{y}}) < G(\boldsymbol{v}^*, \boldsymbol{y}^*)$. We see from (29) that $\Psi(\bar{\boldsymbol{v}}) = G(\bar{\boldsymbol{v}}, \bar{\boldsymbol{y}})$ and $\Psi(\boldsymbol{v}^*) = G(\boldsymbol{v}^*, \boldsymbol{y}^*)$. Hence $\Psi(\bar{\boldsymbol{v}}) < \Psi(\boldsymbol{v}^*)$, which contradicts that $\boldsymbol{v}^*$ is a solution of model (21). This completes the proof. □

A.6. Proof of Proposition 3.7

Proof. Compute that

$$\begin{aligned}
\nabla_{\boldsymbol{v}} H(\boldsymbol{v}, \boldsymbol{y}) &= \boldsymbol{h}_1 + 2\lambda\left(\boldsymbol{h}_2\boldsymbol{h}_2^\top\boldsymbol{v} - \rho\boldsymbol{h}_2\right) + \frac{1}{\gamma}\left(\widetilde{\boldsymbol{I}}^\top\widetilde{\boldsymbol{I}}\boldsymbol{v} - \widetilde{\boldsymbol{I}}^\top\boldsymbol{y}\right), \\
\nabla_{\boldsymbol{y}} H(\boldsymbol{v}, \boldsymbol{y}) &= \frac{1}{\gamma}(\boldsymbol{y} - \widetilde{\boldsymbol{I}}\boldsymbol{v}).
\end{aligned}$$

Let

$$L_1 := 2\lambda\|\boldsymbol{h}_2\|_2^2 + \frac{1}{\gamma} \ \text{ and } \ L_2 := \frac{1}{\gamma}.$$

Then for all $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^{N_1}$ and $\boldsymbol{y} \in \mathbb{R}^N$,

$$\begin{aligned}\|\nabla_{\boldsymbol{v}} H(\boldsymbol{v}_1, \boldsymbol{y}) - \nabla_{\boldsymbol{v}} H(\boldsymbol{v}_2, \boldsymbol{y})\|_2 &= \left\| \left(2\lambda \boldsymbol{h}_2 \boldsymbol{h}_2^\top + \frac{1}{\gamma} \widetilde{\boldsymbol{I}}^\top \widetilde{\boldsymbol{I}}\right)(\boldsymbol{v}_1 - \boldsymbol{v}_2) \right\|_2 \\ &\leqslant \left\| \left(2\lambda \boldsymbol{h}_2 \boldsymbol{h}_2^\top + \frac{1}{\gamma} \widetilde{\boldsymbol{I}}^\top \widetilde{\boldsymbol{I}}\right) \right\|_2 \cdot \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 \\ &\leqslant \left(2\lambda \|\boldsymbol{h}_2\|_2^2 + \frac{1}{\gamma}\right) \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 \\ &= L_1 \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2.\end{aligned}$$

For all $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^N$ and $\boldsymbol{v} \in \mathbb{R}^{N_1}$,

$$\|\nabla_{\boldsymbol{y}} H(\boldsymbol{v}, \boldsymbol{y}_1) - \nabla_{\boldsymbol{y}} H(\boldsymbol{v}, \boldsymbol{y}_2)\|_2 = \frac{1}{\gamma} \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 = L_2 \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2,$$

which completes the proof. □

A.7. Algorithm for ASMCVaR

Algorithm 1 ASMCVaR

Require: Given the sample asset price relative matrix $\boldsymbol{X} \in \mathbb{R}^{T \times N}$; the parameters m and λ; the maximum iteration numbers $(MaxIter_1, MaxIter_2)$ and tolerances (tol_1, tol_2). Set the approximation parameter γ, the expected return level ρ and the confidence level c.

Preparation: Compute the return matrix $\boldsymbol{R} = \boldsymbol{X} - \mathbf{1}_{T \times N}$, the sample mean vector $\hat{\boldsymbol{\mu}} = \frac{1}{T} \boldsymbol{R}^\top \mathbf{1}_T$. Let matrix $\boldsymbol{Q}$ and vector $\boldsymbol{q}$ be given by (16), $\widetilde{\boldsymbol{I}}$ be given by (16), $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ be given by (15). Compute the Lipschitz constants $L_1 = 2\lambda \|\boldsymbol{h}_2\|_2^2 + \frac{1}{\gamma}$ and $L_2 = \frac{1}{\gamma}$.

Initialization: Given the initial vectors by $\boldsymbol{v}^0 = \left(\frac{1}{N}\mathbf{1}_N^\top, \mathbf{0}_{1+T}^\top\right)^\top$ and $\boldsymbol{y}^0 = \frac{1}{N}\mathbf{1}_N$. Set $\beta_1 = \frac{0.99}{L_1}$ and $\beta_2 = \frac{0.99}{L_2}$ in PALM. Set $\theta = \frac{1.99}{\|\boldsymbol{Q}\|_2^2}$ in FPPA.

repeat

$\boldsymbol{p}^k = \boldsymbol{v}^k - \beta_1 \left[\boldsymbol{h}_1 + 2\lambda\left(\boldsymbol{h}_2 \boldsymbol{h}_2^\top \boldsymbol{v}^k - \rho \boldsymbol{h}_2\right) + \frac{1}{\gamma}\left(\widetilde{\boldsymbol{I}}^\top \widetilde{\boldsymbol{I}} \boldsymbol{v}^k - \widetilde{\boldsymbol{I}}^\top \boldsymbol{y}^k\right)\right]$

repeat

1. $\boldsymbol{x}^l = \boldsymbol{Q}\boldsymbol{p}^k + \boldsymbol{z}^l - \theta \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{z}^l$
2. $\boldsymbol{z}^{l+1} = \boldsymbol{x}^l - \max\left\{\boldsymbol{x}^l, \boldsymbol{q}\right\}$
3. $l = l + 1$

until $\frac{\|\boldsymbol{z}^l - \boldsymbol{z}^{l-1}\|_2}{\|\boldsymbol{z}^{l-1}\|_2} \leqslant tol_1$ or $l > MaxIter_1$.

4. $\boldsymbol{v}^{k+1} = \boldsymbol{p}^k - \theta \boldsymbol{Q}^\top \boldsymbol{z}^l$
5. $\boldsymbol{y}^{k+1} = \mathcal{S}_m\left(\boldsymbol{y}^k - \beta_2\left(\frac{1}{\gamma}(\boldsymbol{y}^k - \widetilde{\boldsymbol{I}}\boldsymbol{v}^{k+1})\right)\right)$
6. $k = k + 1$

until $\frac{\|\boldsymbol{v}^k - \boldsymbol{v}^{k-1}\|_2}{\|\boldsymbol{v}^{k-1}\|_2} \leqslant tol_2$ or $k > MaxIter_2$.

Ensure: Portfolio $\boldsymbol{w}^* = \boldsymbol{v}_{1:N}^k$.

A.8. PALM for Relaxed Sparse Regression

To see the performance of the proposed relaxation technique and the PALM algorithm for the sparse regression problem, we consider the following two models:

$$\min_{\boldsymbol{\beta}} \left\{\frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \iota_m(\boldsymbol{\beta})\right\},$$

$$\min_{\boldsymbol{\beta}, \boldsymbol{\eta}} \left\{\frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta} - \boldsymbol{\eta}\|_2^2 + \iota_m(\boldsymbol{\eta})\right\}.$$

They correspond to the original model (19) and the relaxed dual-variable optimization model (24), respectively. We conducted two sets of experiments to validate two aspects:

1. The equivalence of the above relaxation model and original model, in the sense of approximation.
2. The effectiveness of the PALM algorithm for solving the relaxation model.

In these experiments, data sets comprising 50 samples are synthetically generated according to the model $y = \boldsymbol{x}^\top \boldsymbol{\beta}^* + \epsilon$, where $\boldsymbol{\beta}^* = (\mathbf{1}_3, \mathbf{0}_7)^\top \in \mathbb{R}^{10}$ represents the true coefficient vector and ϵ denotes the random error following a normal distribution $\mathcal{N}(0,1)$. The input vector $\boldsymbol{x} \in \mathbb{R}^{10}$ is generated by a normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{10\times 10}$, where $\Sigma_{ij} := 0.5^{|i-j|}$. We set the sparsity level m as 3. The direct exhaustive approach is utilized to enumerate all possible support set configurations, totaling $C_{10}^3 = 120$ cases. By comparing the optimal solutions corresponding to these 120 cases, we can ascertain the exact globally optimal solutions of the aforementioned two models. The iterative scheme of the PALM algorithm for solving the above relaxed model can be given by

$$\begin{cases} \boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k - \alpha_1 \left[\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta}^k - \boldsymbol{y}) + \frac{1}{\gamma}(\boldsymbol{\beta}^k - \boldsymbol{\eta}^k) \right], \\ \boldsymbol{\eta}^{k+1} = \mathcal{S}_m \left(\left(1 - \frac{\alpha_2}{\gamma}\right) \boldsymbol{\eta}^k + \frac{\alpha_2}{\gamma} \boldsymbol{\beta}^{k+1} \right), \end{cases}$$

where $\mathcal{S}_m$ is defined by (27). Parameters α_1 and α_2 are set as $0.99 \cdot \|\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\gamma}\boldsymbol{I}\|_2^{-1}$ and 0.99γ, respectively. To ensure a better convergence, we conduct 5×10^6 iterations of PALM. The initial vectors for iteration are casually chosen as $\boldsymbol{\beta}^0 = \boldsymbol{\eta}^0 = \mathbf{0}_{10}$.

To demonstrate the robustness of our findings, we repeated each of the aforementioned two sets of experiments for 10 times, with different $\boldsymbol{X}$ and $\boldsymbol{y}$ in each run. Through the first set of experiments, we observed that for sufficiently small γ (e.g., 10^{-7}), the solutions of the original and relaxed models are identical. This suggests that as γ tends to zero, the optimal solution of the relaxed model indeed converges to that of the original model. In the subsequent set of experiments, we observed that across the 10 trials, the solutions obtained by PALM consistently approached the optimal solutions of the relaxed model to a remarkable degree. This reflects that in certain scenarios, PALM may be capable of directly converging to the globally optimal solution of the original nonconvex optimization model, even though the initializations may not necessarily be very close to the true solution.

A.9. Cumulative Wealth Plots (Figure 2)

A.10. Transaction Cost Plots (Figure 3)

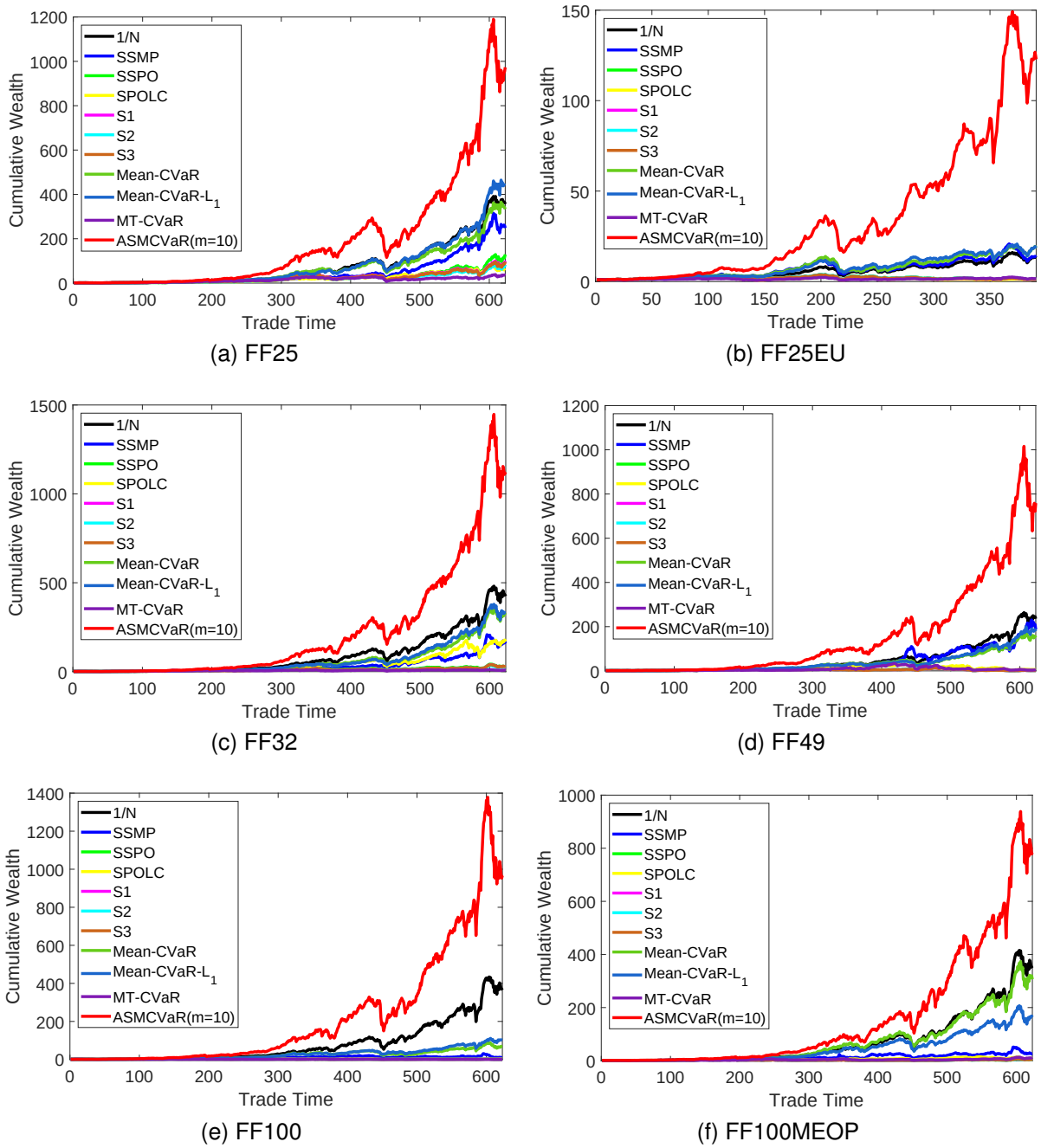

Figure 2. Cumulative wealths of different portfolio optimization models along with trade time on 6 benchmark data sets.

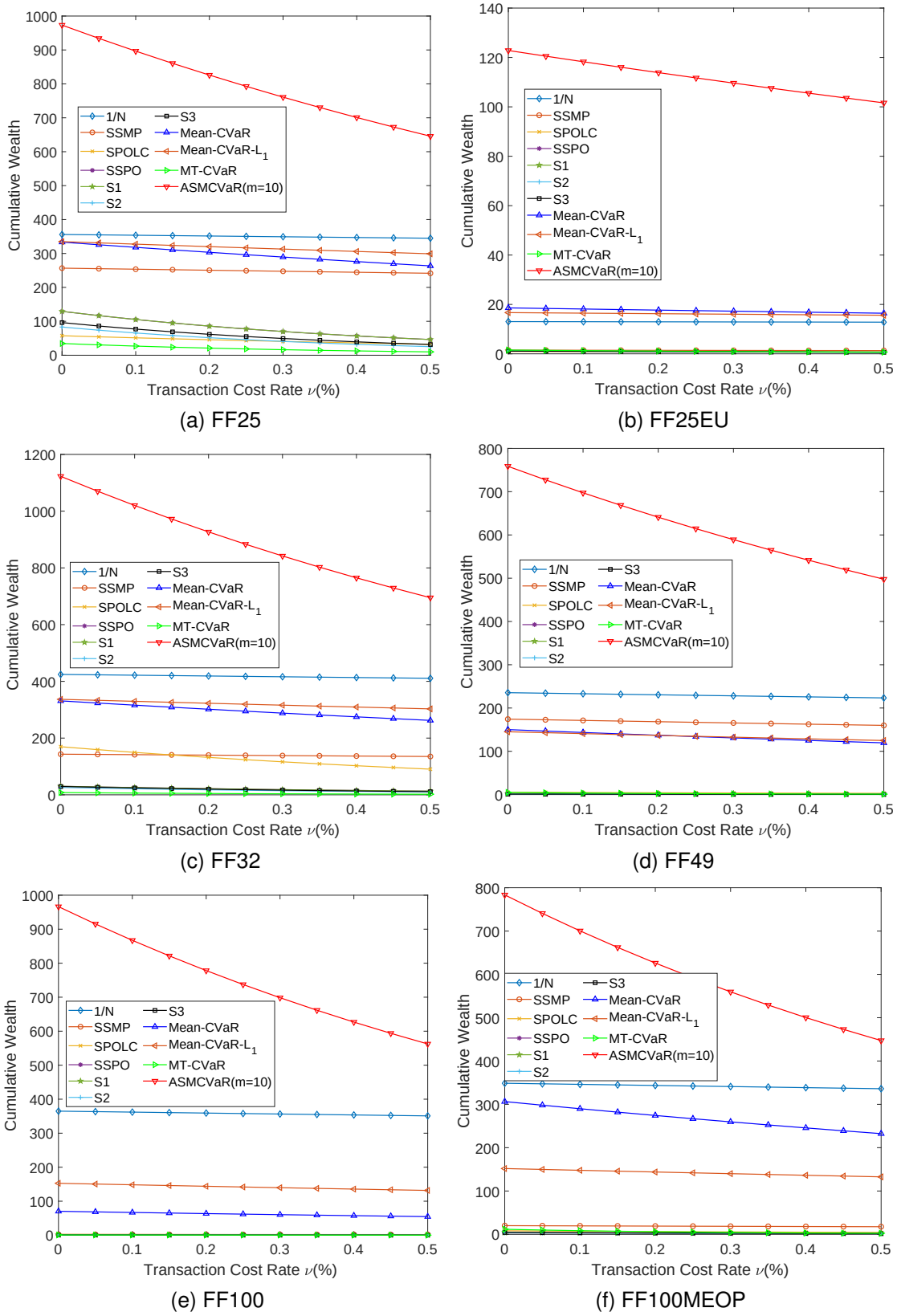

Figure 3. Final cumulative wealths of different portfolio optimization models w.r.t. transaction cost rate ν on 6 benchmark data sets.