Journal of Mathematics and Modeling in Finance (JMMF)
Vol. 5, No. 1, Winter & Spring 2025
Research paper

On data-driven robust portfolio optimization with semi mean absolute deviation via support vector clustering

Eftekhar Kosarinia[1], **Maziar Salahi**[2], **Tahereh Khodamoradi**[3]

[1] Department of Applied Mathematics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran

kosarinia@outlook.com

[2] Department of Applied Mathematics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran

salahim@guilan.ac.ir

[3] Department of Applied Mathematics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran

khodamoradi.tahereh@gmail.com

Abstract:
In [14] the authors have studied robust semi-mean absolute deviation portfolio optimization model when assets expected returns involve uncertainty. They applied a data driven approach via support vector clustering to construct the uncertainty set using support vector clustering. In this paper, we show that their robust formulation is not the worst case counterpart of the original model. Then we give the true robust model of the underlying problems in the best an worst cases. Experiments are conducted to show the optimal objective value of the robust model in [14] belongs to the interval generated by our best and worst case models.

Keywords: Portfolio optimization, Semi-mean absolute deviation, Uncertainty, Support vector clustering.

1 Introduction

Modern portfolio theory or Markowitz mean-variance model, is a mathematical framework for assembling a portfolio of assets such that the expected return is maximized for a given level of risk [9]. However, from practical point of view it ignores many realistic constraints faced by investors. Thus, it was modified to include features like transaction costs [11], multi-period optimization [8], and the cardinality constraint [3]. Also, due to the fact that variance is not a coherence risk measure, other risk measures such as conditional value-at-risk and several others are proposed in the literature [12, 13].

In all of the above-mentioned portfolio optimization models, parameters such as

[2]Corresponding author
Received: 13/03/2025 Accepted: 31/05/2025
https://doi.org/10.22054/JMMF.2025.84881.1170

returns are not known in advance and historical data often are used to predict them. This might lead to inaccurate results thus wrong choice of portfolios. To deal with such uncertainty in portfolio optimization models, a widely used approach is the so called robust optimization which has been applied to several models [1, 4, 7, 10, 15]. An important step in all robust optimization models, is the choice of uncertainty set. Considering uncertainty too big might lead to too conservative solution and considering it too small might left out solutions that might be the the right choice. So choosing it based on the available data come across as a natural choice. Data driven uncertainty set has been variously applied in different optimization models among which is the so called portfolio optimization models [2, 7, 16]. Recently, in [14] the authors have studied portfolio optimization under the semi-mean absolute deviation (SMAD) risk measure [5] under uncertainty and applied the data driven approach of [16] to construct the uncertainty set. They reformulated the robust model as an Linear Programming (LP) problem and performed experiments to show the effectiveness of the robust model compared to the similar models in the literature.

As the authors in [14] treated uncertain return in two constraints differently, thus in this paper we derive the true robust models in the best and worst cases by a different approach. This enables practitioners to obtain portfolio's behavior in the best and worst cases. The derived models are nonlinear compared to the LP model in [14]. Experiments on S&P500 datasets are conducted to confirm our theoretical findings.

The rest of the paper is organized as follows. Section 2 briefly reviews the portfolio optimization problem. Data driven uncertainty set is reviewed in Section 3. In Section 4, we present the best and worst robust counterparts of the underlying portfolio optimization problem under data driven uncertainty set. Finally, experiments are conducted in Section 5.

2 Portfolio optimization model with SMAD risk measure

The SMAD as a risk measure which is proposed by Feinstein and Thapa [5] considers returns below the expected return. It is defined as follows

$$\text{SMAD}(w) = E\left\{|E(R_w) - R_w|_+\right\}.$$

Under the discrete distribution of returns, it becomes

$$\text{SMAD}(w) = \sum_{t=1}^{T} p_t \left(|\sum_{i=1}^{n} r_i w_i - \sum_{i=1}^{n} r_{it} w_i|_+ \right),$$

where $p = (p_1, \ldots, p_T)^T, t = 1, \ldots, T$. Then the portfolio optimization problem that takes a trade-off between SMAD and expected return is as follows:

$$\begin{aligned}
&\min A \\
&\sum_{t=1}^{T} p_t d_t - \sum_{i=1}^{n} r_i w_i \leq A \\
&\sum_{i=1}^{n} r_i w_i - \sum_{i=1}^{n} r_{it} w_i \leq d_t,\ t = 1, \ldots, T \\
&\sum_{i=1}^{n} w_i = 1 \\
&w_i \geq 0,\ i = 1, \ldots, n \\
&d_t \geq 0,\ t = 1, \ldots, T,
\end{aligned} \tag{1}$$

which is an LP problem.

3 Data driven uncertainty set and robust model of [14]

Let $\{r^k\}_{k=1}^N$ be a collection of uncertain returns for n assets i.e., $r^k \in R^n, k = 1, \ldots, N$. In the support vector clustering (SVC) based approach of [16], the goal is to find the smallest ball that includes data points. To avoid large radius ball and exclude outliers, the following soft-margin version is considered:

$$\begin{aligned}
\min_{R,\mathrm{a},\xi} \quad & R^2 + \frac{1}{N\nu} \sum_{k=1}^{N} \xi_k \\
& \|\phi(r^k) - \mathrm{a}\|_2^2 \leq R^2 + \xi_k,\ k = 1, \ldots, N, \\
& \xi_k \geq 0,\ k = 1, \ldots, N,
\end{aligned} \tag{2}$$

where $\phi : R^n \to R^k$ is a kernel function that maps data points to the feature space. For computational efficiency, its dual formulation is used as follows:

$$\begin{aligned}
\min_{\alpha} \quad & \sum_{k=1}^{N} \sum_{k_1=1}^{N} \alpha_k \alpha_{k_1} K(r^k, r^{k_1}) - \sum_{k=1}^{N} \alpha_k K(r^k, r^k) \\
& \alpha_k \leq \frac{1}{N\nu},\ k = 1, \ldots, N \\
& \sum_{k=1}^{N} \alpha_k = 1 \\
& \alpha_k \geq 0,\ k = 1, \ldots, N.
\end{aligned} \tag{3}$$

Shang et al. [16] proposed to use the following weighted generalized kernel:

$$K(r^k, r^{k_1}) = \sum_{i=1}^{n} b_i - \|U(r^k - r^{k_1})\|_1, \tag{4}$$

where $U = \Sigma^{-\frac{1}{2}}$ and Σ is the covariance matrix of the data matrix D. Also, to ensure that kernel is positive definite, b is chosen such that

$$b_i > \max_k u_i^T r^k - \min_k u_i^T r^k,$$

where u_i is the ith column of U. Now let α^* be the optimal solution of the dual model (3). We define the index sets of supprot vector (SV) and boundary SV (BSV) as follows:

$$SV = \{k \mid \alpha_k^* > 0\}, \tag{5}$$

$$BSV = \left\{k \mid 0 < \alpha_k^* < \frac{1}{N\nu}\right\}. \tag{6}$$

Then the radius R can be found by computing the distance between the center a and any boundary support vector $r^{k'}$, $k' \in BSV$ as follows:

$$\begin{aligned}
R^2 &= \|\phi(r^{k'}) - \mathrm{a}\|_2^2 \\
&= \left(\phi(r^{k'}) - \mathrm{a}\right)^T \left(\phi(r^{k'}) - \mathrm{a}\right) \\
&= \phi(r^{k'})^T \phi(r^{k'}) - 2\phi(r^{k'})\mathrm{a} + \mathrm{a}^2 \\
&= \phi(r^{k'})^T \phi(r^{k'}) - 2\alpha_k^* \sum_{k=1}^{N} \phi(r^{k'})\phi(r^k) + \sum_{k=1}^{N}\sum_{k_1=1}^{N} \alpha_k^* \alpha_{k_1}^* \phi(r^k)\phi(r^{k_1}) \\
&= K(r^{k'}, r^{k'}) - 2\sum_{k=1}^{N} \alpha_k^* K(r^{k'}, r^k) + \sum_{k=1}^{N}\sum_{k_1=1}^{N} \alpha_k^* \alpha_{k_1}^* K(r^k, r^{k_1}),\ k' \in BSV.
\end{aligned} \tag{7}$$

Following this, data points r in the uncertainty set $V(D)$ satisfy the following inequality:

$$K(r, r) - 2\sum_{k=1}^{N} \alpha_k^* K(r, r^k) + \sum_{k=1}^{N}\sum_{k_1=1}^{N} \alpha_k^* \alpha_{k_1}^* K(r^k, r^{k_1}) \le R^2. \tag{8}$$

Now using (7), it becomes:

$$K(r, r) - 2\sum_{k=1}^{N} \alpha_k^* K(r, r^k) + \sum_{k=1}^{N}\sum_{k_1=1}^{N} \alpha_k^* \alpha_{k_1}^* K(r^k, r^{k_1}) \le$$

$$K(r^{k'}, r^{k'}) - 2\sum_{k=1}^{N} \alpha_k^* K(r^{k'}, r^k) + \sum_{k=1}^{N}\sum_{k_1=1}^{N} \alpha_k^* \alpha_{k_1}^* K(r^k, r^{k_1}),$$

$$-2\sum_{k=1}^{N}\alpha_k^* K(r,r^k) \le -2\sum_{k=1}^{N}\alpha_k^* K(r^{k'},r^k),$$

$$\sum_{k=1}^{N}\alpha_k^* K(r,r^k) \ge \sum_{k=1}^{N}\alpha_k^* K(r^{k'},r^k),$$

$$V(D)=\left\{r \mid \sum_{k=1}^{N}\alpha_k^* K(r,r^k) \ge \sum_{k=1}^{N}\alpha_k^* K(r^{k'},r^k),\ k'\in BSV\right\}.$$

Since for $\alpha_k^*=0, k\notin SV$ and using (4) we further have

$$\sum_{k\in SV}\alpha_k^*\left(\sum_{i=1}^{n}b_i-\|U(r-r^k)\|_1\right) \ge \sum_{k\in SV}\alpha_k^*\left(\sum_{i=1}^{n}b_i-\|U(r^{k'}-r^k)\|_1\right),$$

$$\sum_{k\in SV}\sum_{i=1}^{n}\alpha_k^* b_i-\sum_{k\in SV}\alpha_k^*\|U(r-r^k)\|_1 \ge \sum_{k\in SV}\sum_{i=1}^{n}\alpha_k^* b_i-\sum_{k\in SV}\alpha_k^*\|U(r^{k'}-r^k)\|_1,$$

$$\sum_{k\in SV}\alpha_k^*\|U(r-r^k)\|_1 \le \sum_{k\in SV}\alpha_k^*\|U(r^{k'}-r^k)\|_1,$$

thus

$$V(D)=\left\{r \mid \sum_{k\in SV}\alpha_k^*\|U(r-r^k)\|_1 \le \sum_{k\in SV}\alpha_k^*\|U(r^{k'}-r^k)\|_1,\ k'\in BSV\right\}.$$

Finally, taking

$$\theta=\min_{k'\in BSV}\left\{\sum_{k\in SV}\alpha_k^*\|U(r^{k'}-r^k)\|_1\right\}$$

and

$$z_k=|U(r-r^k)|\in\mathbb{R}^n,\ k\in SV,$$

we have

$$V(D)=\left\{r \mid \exists z_k\ s.t.\ \sum_{k\in SV}\alpha_k^* z_k^T e\le\theta \text{ and } -z_k\le U(r-r^k)\le z_k,\ k\in SV\right\}, \quad (9)$$

where $e=(1,\ldots,1)^T\in\mathbb{R}^n$.

The proposed robust portfolio optimization model in [14] using SMAD as a risk

measure under the uncertainty set $V(D)$ is as follows:

$$\begin{aligned}
\min \quad & A \\
& \sum_{t=1}^{T} p_t d_t + \max_{r\in V(D)} (-r^T w) \le A, \\
& \max_{r\in V(D)} (r^T w) - \sum_{i=1}^{n} r_{it} w_i - d_t \le 0,\ t = 1, \dots, T, \\
& \sum_{i=1}^{n} w_i = 1, \\
& w_i \ge 0,\ i = 1, \dots, n, \\
& d_t \ge 0,\ t = 1, \dots, T.
\end{aligned} \tag{10}$$

Using LP duality for $\max_{r\in V(D)}(-r^T w)$ and $\max_{r\in V(D)}(r^T w)$, the authors in [14] have obtained the following equivalent model of (10):

$$\begin{aligned}
\min \quad & A \\
& \sum_{t=1}^{T} p_t d_t + \sum_{k\in SV} (\mu_k - \lambda_k)^T U r^k + \theta\eta \le A, \\
& \sum_{k\in SV} U(\lambda_k - \mu_k) - w = 0, \\
& \mu_k + \lambda_k = \eta.\alpha_k.e,\ k \in SV, \\
& \sum_{k\in SV} (\xi_k - \tau_k)^T U r^k + \theta\delta - \sum_{i=1}^{n} r_{it} w_i - d_t \le 0,\ t = 1, \dots, T, \\
& \sum_{k\in SV} U(\tau_k - \xi_k) + w = 0, \\
& \xi_k + \tau_k = \delta\alpha_k e,\ k \in SV, \\
& \sum_{i=1}^{n} w_i = 1, \\
& w_i \ge 0,\ i = 1, \dots, n, \\
& \eta, \delta \ge 0,\ \mu_k, \lambda_k, \tau_k, \xi_k \in \mathbb{R}^n_+,\ k \in SV, \\
& d_t \ge 0,\ t = 1, \dots, T.
\end{aligned} \tag{11}$$

As we see, the authors in [14] have treated uncertain r in first and second set of constraints of model (10) differently, thus the two maximums might attain at two different $r \in V(D)$. Therefore it does not give the robust model of (1) in the worst case. Following this weakness, in the next section we provide the robust counterparts of (1) in the best and worst cases.

4 The best and worst robust models

In this section, we give the robust version of (1) in the best and worst cases. This gives us an interval that contains the optimal objective value of model (11).

4.1 Best case

The best case (lower bound) of model (1) when r is uncertain and belong to the set $\bar{V}(D)$ is the solution of the following problem:

$$\min_{r\in\bar{V}(D)} \min_{w} \left(\sum_{t=1}^{T} p_t |r^T w - \sum_{i=1}^{n} r_{it} w_i|_+ - r^T w \right), \tag{12}$$

which can be written as follows

$$\begin{aligned} \min \quad & \sum_{t=1}^{T} p_t d_t - \sum_{i=1}^{n} r_i w_i \\ & \sum_{i=1}^{n} r_i w_i - \sum_{i=1}^{n} r_{it} w_i \le d_t, t = 1, \dots, T \\ & \sum_{i=1}^{n} w_i = 1 \\ & -z_k \le U(r - r^k) \le z_k, k \in SV \\ & w_i \ge 0, i = 1, \dots, n \\ & d_t \ge 0, t = 1, \dots, T. \end{aligned} \tag{13}$$

One can see that the first term in the first set of constraint is bilinear, thus the problem becomes nonlinear and nonconvex as opposed to the LP model in [14].

4.2 Worst case

The worst case (upper bound) of model (1) under the uncertainty $V(D)$ is as follows:

$$\max_{r\in V(D)} \min_{w} \left(\sum_{t=1}^{T} p_t |r^T w - \sum_{i=1}^{n} r_{it} w_i|_+ - r^T w \right). \tag{14}$$

First we focus on the inner minimization problem. Let $d_t = |\sum_{i=1}^n r_i w_i - \sum_{i=1}^n r_{it} w_i|_+$, then the minimization problem becomes

$$\begin{aligned} \min \quad & \sum_{t=1}^T p_t d_t - \sum_{i=1}^n r_i w_i \\ & \sum_{i=1}^n r_i w_i - \sum_{i=1}^n r_{it} w_i \le d_t, \ t = 1, \dots, T \\ & \sum_{i=1}^n w_i = 1 \\ & w_i \ge 0, \ i = 1, \dots, n \\ & d_t \ge 0, \ t = 1, \dots, T, \end{aligned}$$

and its dual is

$$\begin{aligned} \max \quad & \phi \\ & \sum_{t=1}^T r_i \tau_t - \sum_{t=1}^T r_{it} \tau_t + \phi \le -r_i, \ i = 1, \dots, n \\ & -\tau_t \le p_t, \ t = 1, \dots, T, \\ & -\tau_t \ge 0, \ t = 1, \dots, T. \end{aligned}$$

Then using strong duality in LP, (14) becomes

$$\begin{aligned} \max \quad & \phi \\ & \sum_{t=1}^T r_i \tau_t - \sum_{t=1}^T r_{it} \tau_t + \phi \le -r_i, \ i = 1, \dots, n \\ & -\tau_t \le p_t, \ t = 1, \dots, T \\ & \sum_{k \in SV} \alpha_k z_k^T e \le \theta \\ & -z_k \le U(r - r^k) \le z_k, \ k \in SV \\ & -\tau_t \ge 0, \ t = 1, \dots, T. \end{aligned} \tag{15}$$

Similar to the best case, the first term in the first set of constraints are bilinear thus the problem becomes nonconvex.

Lemma 4.1. *Let OPT_{LB} and OPT_{UB} be the optimal objective values of models (13) and (15). Then the interval $[OPT_{LB}, OPT_{UB}]$ contains OPT_R, the optimal objective value of model (11).*

Proof: It easily follows from the construction of models (13) and (15). □

5 Experimental results

In this section, we conducted experiments on the datasets of 50 assets from S& P 500 to validate the theoretical results numerically as well. All implementations are done in MATLAB, CVX [6] is used to solve convex optimization models (1) and (11) and 'fmincon' command of MATLAB is used to solve nonlinear and nonconvex models (13) and (15). The results are summarized in Tables 3-8, where we report optimal objective values of all models, risks and mean returns. Portfolio's return is computed as $y_t = \sum_{i=1}^{n} r_{it} w_i^*, t = 1, \ldots, T$, portfolio's mean return, and standard deviation also are computed as follows:

$$\bar{y} = \frac{\sum_{t=1}^{T} y_t}{T}, \sigma = \sqrt{\frac{\sum_{t=1}^{T} (\bar{y} - y_t)^2}{T-1}}.$$

As we see in Table 3, for all scenarios the objective function of model (11) lies in between objective values of models (13) and (15) as also proved in Lemma 4.1. In Tables 4-8 for different scenarios, we report the risks, returns and their ratios. As we see, both best and worst case models have better return to risk ratios compared to models (1) and (11). These results confirm our theoretical development.

Table 1: Objective function values for $n = 50$, $T = 24$.

N	Model (1)	Model (11)	Model (13)	Model (15)
$N = 100$	-0.0261	0.0134	0.0117	0.0162
$N = 150$	-0.0261	0.0160	0.0148	0.0193
$N = 200$	-0.0261	0.0152	0.0135	0.0185
$N = 250$	-0.0261	0.0178	0.0156	0.0200
$N = 300$	-0.0261	0.0183	0.0166	0.0215

Table 2: Returns, risks and their ratios for $n = 50$, $T = 24$, $N = 100$.

	Model (1)	Model (11)	Model (13)	Model (15)
Return	0.0395	0.0174	0.0370	0.0381
Risk	0.0399	0.0218	0.0358	0.0362
$\frac{\text{Return}}{\text{Risk}}$	0.9883	0.7982	1.0335	1.0532

Table 3: Returns, risks and their ratios for $n = 50$, $T = 24$, $N = 150$.

	Model (1)	Model (11)	Model (13)	Model (15)
Return	0.0395	0.0170	0.0357	0.0372
Risk	0.0399	0.0219	0.0353	0.0359
$\frac{\text{Return}}{\text{Risk}}$	0.9883	0.7735	1.0124	1.0354

Table 4: Returns, risks and their ratios for $n = 50$, $T = 24$, $N = 200$.

	Model (1)	Model (11)	Model (13)	Model (15)
Return	0.0395	0.0177	0.0360	0.0398
Risk	0.0399	0.0222	0.0351	0.0381
$\frac{\text{Return}}{\text{Risk}}$	0.9883	0.7975	1.0251	1.0426

Table 5: Returns, risks and their ratios for $n = 50$, $T = 24$, $N = 250$.

	Model (1)	Model (11)	Model (13)	Model (15)
Return	0.0395	0.0164	0.0328	0.0386
Risk	0.0399	0.0228	0.0326	0.0379
$\frac{\text{Return}}{\text{Risk}}$	0.9883	0.7207	1.0048	1.0183

Table 6: Returns, risks and their ratios for $n = 50$, $T = 24$, $N = 300$.

	Model (1)	Model (11)	Model (13)	Model (15)
Return	0.0395	0.0167	0.0374	0.0389
Risk	0.0399	0.0207	0.0357	0.0364
$\frac{\text{Return}}{\text{Risk}}$	0.9883	0.8094	1.0473	1.0691

Bibliography

[1] D. BERTSIMAS, M. SIM, *The price of robustness*, Operations Research, 52(2004), pp. 3553.

[2] D. BERTSIMAS, V. GUPTA, N. KALLUS, *Data-driven robust optimization*, Mathematical Programming, 167(2018), pp. 235292.

[3] T.J. CHANG, N. MEADE, J.E. BEASLEY, Y.M. SHARAIHA, *Heuristics for cardinality constrained portfolio optimisation*. Computers & Operations Research 27(13)(2000), pp. 12711302.

[4] L. EL GHAOUI, M. OKS, F. OUSTRY, *Meshless methods*, Worst-case value-at-risk and robust portfolio optimization: A conic programming approach, Operations Research, 51(2003), pp. 543556.

[5] C. D. FEINSTEIN, M. N. THAPA, *A reformulation of a mean-absolute deviation portfolio optimization model*, Management Science, 39 (1993), pp. 1552-1553.

[6] M. Grant, S. Boyd, Y. Ye, *CVX: Matlab software for disciplined convex programming, version 2.0 beta*, 2013.

[7] R. Ji, M. A. Lejeune, *Data-driven optimization of reward-risk ratio measures*, INFORMS Journal on Computing, 33 (2021), pp. 11201137.

[8] D. Li, W.L. Ng *Optimal dynamic portfolio selection: Multiperiod mean-variance formulation*, Mathematical Finance 10(3)(2000), pp. 387406.

[9] H. Markowitz, *Portfolio selection*, Journal of Finance, 7(1) (1952), pp. 7791.

[10] Y. Moon, T. Yao, *A robust mean absolute deviation model for portfolio optimization*, Computers and Operations Research, 38(2011), pp. 1251-1258.

[11] PN.R. Patel, M.G. Subrahmanyam *A simple algorithm for optimal portfolio selection with fixed transaction costs*, Management Science 28(3)91982), pp. 303314.

[12] R.T. Rockafellar, S. Uryasev, *Optimization of conditional value-at-risk*, The Journal of Risk, 2 (2000), pp. 21–42.

[13] M. Salahi, T. Khodamoradi, and A. Hamdi, *Mean-standard deviation-conditional value-at-risk portfolio optimization*, Journal of Mathematics and Modeling in Finance, 3(1) (2023), pp. 83-98.

[14] R. Sehgal, P. Jagadeshm, *Data-driven robust portfolio optimization with semi mean absolute deviation via support vector clustering*, Expert Systems with Applicatins, 224(2023), 120000.

[15] R. Sehgal, A. Mehra, *Robust rewardrisk ratio portfolio optimization*, International Transactions in Operational Research, 28 (2021a), pp. 21692190.

[16] C. Shang, X. Huang, F. You, *Data-driven robust optimization based on kernel learning*, Computers and Chemical Engineering, 106 (2017), pp. 464479.

How to Cite: Eftekhar Kosarinia[1], Maziar Salahi[2], Tahereh Khodamoradi[3], *On data-driven robust portfolio optimization with semi mean absolute deviation via support vector clustering*, Journal of Mathematics and Modeling in Finance (JMMF), Vol. 5, No. 1, Pages:155–165, (2025).