Comput Optim Appl (2015) 62:613–639
DOI 10.1007/s10589-015-9755-3

Optimizing over coherent risk measures and non-convexities: a robust mixed integer optimization approach

Dimitris Bertsimas[1] · **Akiko Takeda**[2]

Received: 7 July 2014 / Published online: 3 May 2015
© Springer Science+Business Media New York 2015

Abstract Recently, coherent risk measure minimization was formulated as robust optimization and the correspondence between coherent risk measures and uncertainty sets of robust optimization was investigated. We study minimizing coherent risk measures under a norm equality constraint with the use of robust optimization formulation. Not only existing coherent risk measures but also a new coherent risk measure is investigated by setting a new uncertainty set. The norm equality constraint itself has a practical meaning or plays a role to prevent a meaningless solution, the zero vector, in the context of portfolio optimization or binary classification in machine learning, respectively. For such advantages, the convexity is sacrificed in the formulation. However, we show a condition for an input of our problem which guarantees that the nonconvex constraint is convexified without changing the optimality of the problem. If the input does not satisfy the condition, we propose to solve a mixed integer optimization problem by using the ℓ_1 or ℓ_∞-norm. The numerical experiments show that our approach has good performance for portfolio optimization and binary classification and also imply its flexibility of modelling that makes it possible to deal with various coherent risk measures.

Keywords Coherent risk measure minimization · Robust optimization · Nonconvexity · Portfolio optimization · Binary classification

✉ Akiko Takeda
takeda@mist.i.u-tokyo.ac.jp

Dimitris Bertsimas
dbertsim@mit.edu

[1] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2] Department of Mathematical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

1 Introduction

Uncertainty is an inevitable feature of many decision-making environments. Managers, professionals, and others need to make decisions to optimize a system with incomplete information and considerable uncertainty. To formulate optimization problems that are defined by uncertain inputs, it is important to define a risk that the concerned system has. One often regards uncertain inputs of the system as random variables and uses the expected value or the variance of uncertain costs involved in a decision as a risk measure.

Artzner et al. [1] presented an axiomatic definition of risk measures satisfying four natural properties and termed such risk measures *coherent*. The coherent risk measure has reasonable properties such as "sub-additivity", which implies that diversification leads to less risk. A popular example of a coherent risk measure is *conditional value-at-risk* (CVaR). Rockafellar and Uryasev [13,14] proposed to minimize the CVaR for optimizing a portfolio so as to reduce the risk of high losses. Ruszczynski and Shapiro [15] considered optimization problems involving convex risk measures (which include coherent risk measures) and developed theoretical works on the optimality and duality theorem for those problems.

Robust optimization (RO) is another approach for optimization under uncertainty. The objective of robust optimization models and algorithms is to obtain solutions that are guaranteed to perform well (in terms of feasibility or near-optimality) for all possible realizations of the uncertain input parameters. The range of possible realizations is given as a set $\mathcal{U}$ called *uncertainty set*. Recently, Bertsimas and Brown [2] and Natarajan et al. [11] independently formulated coherent risk measure minimization as robust optimization and showed the correspondence between coherent risk measures and uncertainty sets $\mathcal{U}$.

In this paper, we consider minimizing a coherent risk measure under a norm equality constraint with the use of robust optimization formulation. Not only well-known coherent risk measures but also a new coherent risk measure is investigated by setting a new uncertainty set $\mathcal{U}$ in numerical experiments. Concretely, a coherent risk measure is minimized with respect to a decision variable $\boldsymbol{v}$ under a norm equality constraint (e.g., $\|\boldsymbol{v}\| = c$ for a positive value c). The concerned uncertain optimization problems stem from classification in machine learning and portfolio optimization in finance. The norm equality constraint itself has a practical meaning in the context of portfolio optimization and it also plays a role to prevent the meaningless solution $\boldsymbol{v} = \mathbf{0}$ in the context of classification in machine learning. For such advantages, the convexity is sacrificed in the formulation. However, we show a condition of c which guarantees that $\|\boldsymbol{v}\| = c$ is convexified to $\|\boldsymbol{v}\| \leq c$ without changing the optimality of the problem. If c satisfies the condition, we can solve a convex problem with $\|\boldsymbol{v}\| \leq c$ for the original problem.

When c does not satisfy the condition, a nonconvex optimization problem including $\|\boldsymbol{v}\| = c$ has to be solved. We reformulate a coherent risk minimization problem including an ℓ_1 or ℓ_∞-norm constraint as a mixed integer optimization (MIO) problem.

In portfolio optimization, the norm-constraint model has recently been studied by various researchers. DeMiguel et al. [7] used variance as the risk measure in the norm-constrained portfolio optimizations with various types of norms (note that variance is

not coherent), while Gotoh and Takeda [9] used CVaR risk measure. Brodie et al. [5] incorporated an ℓ_1-norm penalty on the portfolio decision vector into the traditional Markowitz portfolio optimization model in order to encourage sparse portfolios. Our portfolio optimization model which minimizes a coherent risk measure under an ℓ_p-norm constraint can be regarded as a general model, compared with these existing portfolio models.

As a common source of nonconvexity in practical portfolio optimization problems, Stubbs and Vandenbussche [18] have referred to leverage requirement in addition to threshold constraints on the holdings or trades. Though they recommended that such constraints be left out of analyses because there is no theory to support the required optimality conditions, we can obtain a global optimal solution for nonconvex portfolio optimization including leverage requirement by reformulating it as an MIO formulation.

For classification in machine learning, loss functions (measures of misclassification) to be minimized were recently related to financial risk measures. For example, Xu et al. [22] proposed a comprehensive robust classification model that uses a discounted loss function depending on realized data and investigated the relationship between comprehensive robustness and convex risk measures. Takeda and Sugiyama [19] showed that binary classification models such as ν-SVM [17] and its extended model, Eν-SVM [12], minimize the CVaR of margin distribution. ν-SVM has the ℓ_2-norm constraint $\|\boldsymbol{v}\|_2 \leq c$, whereas E$\nu$-SVM has the nonconvex one $\|\boldsymbol{v}\|_2 = c$. Goto et al. [10] extended the CVaR-based classification methods to those based on coherent risk measures such as the mean absolute semi-deviation (MASD). They used the ℓ_2 norm for $\|\boldsymbol{v}\|_p$ and proposed a local solution method for the resulting problem including $\|\boldsymbol{v}\|_2 = c$. If we use the ℓ_1 or ℓ_∞ norm for $\|\boldsymbol{v}\|_p$, our MIO formulation gives a global optimal solution for the classification model minimizing a coherent risk measure. The ℓ_1-norm constraint also contributes to feature selection that analyzes the impact of its features on the model.

We report computational results in the context of portfolio optimization and machine learning that demonstrate that

(a) the approach leads to a new coherent risk minimization model by preparing a new set $\mathcal{U}$ with the use of, e.g., a Bregman divergence;
(b) it leads to improved out-of-sample performance in a machine learning context, and moreover, in the convex case where $\|\boldsymbol{v}\|_1 \leq c$ or $\|\boldsymbol{v}\|_\infty \leq c$, the computational time is smaller than existing machine learning approach using the ℓ_2-norm constraint;
(c) it shows flexibility of the proposed approach to model side constraints as well as integrality considerations and leads to improved out-of-sample performance in a portfolio optimization context.

The paper is organized as follows. Section 2 presents robust optimization formulation for coherent risk minimization under a norm equality constraint. Section 3 shows a condition for reducing the optimization problem to a convex problem without changing the optimality of the problem. For the cases where the condition is not satisfied, Sect. 4 provides MIO formulations using the ℓ_1 or ℓ_∞ norm for the norm constraint. The coherent risk minimization under a norm equality constraint is used for portfolio optimization or binary classification in Sect. 5 and the performance of our model is

compared against those of popular portfolio models and machine learning models in Sect. 6. Sect. 7 concludes the paper.

2 Robust optimization formulation for coherent risk minimization

2.1 Coherent risk measure

Consider a random variable $\tilde{\boldsymbol{z}}$ in $\mathbb{R}^n$. $\tilde{\boldsymbol{z}}$ could denote the returns of the assets in portfolio optimization, and it also denotes the feature vector for classification in machine learning. We restrict our attention to the space $\mathcal{V}$ defined as an affine combination of the random variables $\tilde{\boldsymbol{z}}$ by following [2,11]:

$$\mathcal{V} := \left\{ \tilde{v} : \exists \boldsymbol{v} \text{ such that } \tilde{v} = \tilde{\boldsymbol{z}}^\top \boldsymbol{v} \right\}.$$

Definition 1 [1] A function $\mu : \mathcal{V} \to \mathbb{R}$ that satisfies the following four axioms for all random variables $\tilde{v}, \tilde{w} \in \mathcal{V}$ is called a *coherent risk measure*.

- monotonicity: if $\tilde{v} \geq \tilde{w}$, then $\mu(\tilde{v}) \leq \mu(\tilde{w})$.
- translation invariance: if $a \in \mathbb{R}$, then $\mu(\tilde{v} + a) = \mu(\tilde{v}) - a$.
- subadditivity: $\mu(\tilde{v} + \tilde{w}) \leq \mu(\tilde{v}) + \mu(\tilde{w})$.
- positive homogeneity: If $\lambda \geq 0$, then $\mu(\lambda \tilde{v}) = \lambda \mu(\tilde{v})$.

2.2 Coherent risk measure minimization

For a coherent risk measure μ and a random variable $\tilde{\boldsymbol{z}}$ in $\mathbb{R}^n$, we consider optimizing over coherent risk measures:

$$\min_{\|\boldsymbol{v}_s\|_p = c, \boldsymbol{v} \in V} \mu\left(\tilde{\boldsymbol{z}}^\top \boldsymbol{v}\right), \tag{1}$$

where V is a convex set. Let $\mathcal{S}$ be a subset of the indices $\{1, \ldots, n\}$ of variables $\boldsymbol{v} \in \mathbb{R}^n$ and $\boldsymbol{v}_s \in \mathbb{R}^{n'}$ $(0 < n' \leq n)$ denotes a subvector of $\boldsymbol{v} \in \mathbb{R}^n$ corresponding to the indices of $\mathcal{S}$. $\|\cdot\|_p$ denotes the ℓ_p norm $(p \geq 1)$. Let a constant c be positive to avoid the case of $\boldsymbol{v}_s = \boldsymbol{0}$ and, moreover, we assume the following condition for c.

Assumption 1 c (> 0) of (1) satisfies

$$\min_{\boldsymbol{v} \in V} \|\boldsymbol{v}_s\|_p \leq c \leq \max_{\boldsymbol{v} \in V} \|\boldsymbol{v}_s\|_p$$

so that (1) is feasible.

It is well-known that any coherent risk measure can be equivalently described in terms of the worst-case expectation over a family of distributions $\mathcal{Q}$, and therefore,

coherent risk minimization problem (1) can be written as follows (see, e.g., representation theorem for coherent risk measures in [1]):

$$\min_{\|\boldsymbol{v}_s\|_p=c, \boldsymbol{v}\in V} \max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}_{\boldsymbol{q}}\left(-\tilde{\boldsymbol{z}}^\top \boldsymbol{v}\right), \tag{2}$$

where $\mathbb{E}_{\boldsymbol{q}}(\tilde{u})$ denotes the expectation of the random variable $\tilde{u}$ under $\boldsymbol{q}$. (2) is furthermore equivalent to the robust optimization problem with some convex set $\mathcal{U}$ corresponding to $\mathcal{Q}$:

$$\min_{\|\boldsymbol{v}_s\|_p=c, \boldsymbol{v}\in V} \max_{z\in\mathcal{U}} \; -\boldsymbol{z}^\top \boldsymbol{v}. \tag{3}$$

in both cases of the finite probability space for $\tilde{\boldsymbol{z}}$ (see [2]) and more general probability space for $\tilde{\boldsymbol{z}}$ (see [11]). For simplicity of the discussion, we assume $\mathcal{U}$ is bounded so that (3) has an optimal solution under Assumption 1.

Here we assume finite probability space for $\tilde{\boldsymbol{z}}$ by following [2]. Denote the support of $\tilde{\boldsymbol{z}}$ by $\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$ and define the matrix form $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m]$. We assume $\hat{p}_i = P(\tilde{\boldsymbol{z}} = \boldsymbol{z}_i)$ as a reference probability. Note that $\hat{\boldsymbol{p}} \in \Delta^m$, where $\Delta^m = \{\boldsymbol{q} \in \mathbb{R}^m : \boldsymbol{e}^\top \boldsymbol{q} = 1,\ \boldsymbol{q} \geq \boldsymbol{0}\}$ and $\boldsymbol{e}$ is the all-one vector. Theorem 3.1 of [2] shows a relation between $\mathcal{Q}$ of (2) and $\mathcal{U}$ of (3) as

$$\begin{aligned} &\mathcal{U} = \text{conv}\left(\{\boldsymbol{Z}\boldsymbol{q} : \boldsymbol{q} \in \mathcal{Q}\}\right), \text{ or} \\ &\mathcal{Q} = \left\{\boldsymbol{q} \in \Delta^m : \ \boldsymbol{Z}\boldsymbol{q} \in \mathcal{U}\right\} \ \text{ if } \mathcal{U} \subseteq \text{conv}(\mathcal{Z}), \end{aligned} \tag{4}$$

where conv means the convex hull of a set. The main interest of this paper is in deriving coherent risk optimization models (3) for finance or machine learning applications by choosing $\mathcal{U}$ appropriately. We will show later examples of $\mathcal{U}$ and how to deal with the nonconvex constraint $\|\boldsymbol{v}_s\|_p = c$ in (3).

3 Condition for convexifying the norm constraint

Equation (3) is a nonconvex optimization problem due to the norm constraint $\|\boldsymbol{v}_s\|_p = c$. However, we can replace the nonconvex constraint by a convex one $\|\boldsymbol{v}_s\|_p \leq c$ without changing the optimality if (3) satisfies a condition. We investigate the condition in this section.

3.1 Arbitrary convex set for V

Here we assume an arbitrary convex set for V in (3). Note that $\mathcal{U}$ constructed by (4) is a closed convex set. We define a convex optimization problem constructed by removing the nonconvex constraint $\|\boldsymbol{v}_s\|_p = c$ from (3),

$$\inf_{\boldsymbol{v}\in V} \max_{z\in\mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v}, \tag{5}$$

and let $\hat{\boldsymbol{v}}$ and $\hat{f}$ be, respectively, an optimal solution and the optimal value of (5) if they exist. Then we define the norm-threshold by

$$\tau \equiv \|\hat{\boldsymbol{v}}_s\|_p. \tag{6}$$

If (5) is unbounded, let $\hat{f} = -\infty$ and $\tau = \infty$. We use the norm-threshold τ for reducing (3) to the convex relaxation problem as the following theorem shows.

Theorem 1 *Suppose that the parameter c of (3) satisfies $c \leq \tau$. If the optimal value of the convex problem*

$$\min_{\|\boldsymbol{v}_s\|_p \leq c, \boldsymbol{v} \in V} \max_{z \in \mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v} \tag{7}$$

is strictly larger than $\hat{f}$ of (5), the optimal solution of (7) is also optimal to (3).

Proof Let $\boldsymbol{v}^*$ be an optimal solution of (7) and remind that $\hat{\boldsymbol{v}}$ and $\hat{f}$ are respectively an optimal solution and the optimal value of (5) when (5) is bounded (we will consider the unbounded case later). By the assumption of this theorem, the optimal value of (7) is larger than $\hat{f}$, that is,

$$\hat{f} = \max_{z \in \mathcal{U}} \left(-\boldsymbol{z}^\top \hat{\boldsymbol{v}}\right) < \max_{z \in \mathcal{U}} \left(-\boldsymbol{z}^\top \boldsymbol{v}^*\right). \tag{8}$$

Equation (7) is constructed by adding the constraint $\|\boldsymbol{v}_s\|_p \leq c$ to (5), and therefore, the optimal value of (7) is larger than or equal to $\hat{f}$. The assumption deletes the possibility where two optimization problems have the same optimal value. When (5) is unbounded, let $\hat{\boldsymbol{v}}$ be a feasible solution of (5) satisfying (8).

To prove this theorem, it is sufficient to show $\|\boldsymbol{v}_s^*\|_p = c$. Here we assume $\|\boldsymbol{v}_s^*\|_p < c$ on contrary. Then, because of $\|\hat{\boldsymbol{v}}_s\|_p = \tau \geq c$, we obtain

$$\boldsymbol{v}(\tilde{\lambda}) \equiv \tilde{\lambda} \hat{\boldsymbol{v}} + \left(1 - \tilde{\lambda}\right) \boldsymbol{v}^*$$

satisfying $\|\boldsymbol{v}(\tilde{\lambda})_s\|_p = c$ with some $\tilde{\lambda} \in (0, 1]$. Note that $\boldsymbol{v}(\tilde{\lambda})$ is a feasible solution of (3) and (7). Moreover, we have

$$\max_{z \in \mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v}(\tilde{\lambda}) \leq \tilde{\lambda} \max_{z \in \mathcal{U}} \left(-\boldsymbol{z}^\top \hat{\boldsymbol{v}}\right) + \left(1 - \tilde{\lambda}\right) \max_{z \in \mathcal{U}} \left(-\boldsymbol{z}^\top \boldsymbol{v}^*\right) < \max_{z \in \mathcal{U}} \left(-\boldsymbol{z}^\top \boldsymbol{v}^*\right),$$

where the strict inequality comes from (8). $\boldsymbol{v}(\tilde{\lambda})$ is a better solution than $\boldsymbol{v}^*$ to (7), which contradicts the optimality of $\boldsymbol{v}^*$. □

For (3) with $c > \tau$, the convex relaxation problem (7) will give $\hat{\boldsymbol{v}}$ as its optimal solution because of $\tau = \|\hat{\boldsymbol{v}}_s\|_p < c$. Therefore, (7) can not give an optimal solution of (3).

3.2 Convex cone $\mathcal{K}$ for V

For V of (3), we assume a special set, convex cone $\mathcal{K}$, which includes $\mathcal{K} = \mathbb{R}^n$ of binary classification methods (see Sect. 5.2) and $\mathcal{K} = \mathbb{R}^n_+$ of long-only portfolio optimization (see Sect. 5.1.2). In the case of $V = \mathcal{K}$, as shown later in Remark 1, the threshold τ has only two possible values (0 or ∞) when the optimal value (3) is nonzero. We can furthermore relate the transferability test of (3), $c \leq \tau$, to a geometric condition for $\mathcal{U}$. As a result, instead of computing τ by (6), we can use another criterion such as the relative position of $\mathbf{0}$ to $\mathcal{U}$ in order to check the transferability of (3) to the convex relaxation problem (7).

We deal with the following problem

$$\min_{\|\boldsymbol{v}_s\|_p = c, \boldsymbol{v} \in \mathcal{K}} \max_{\boldsymbol{z} \in \mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v} \quad \left(\text{or} \quad \min_{\|\boldsymbol{v}_s\|_p = c, \boldsymbol{v} \in \mathcal{K}} \mu \left(\tilde{\boldsymbol{z}}^\top \boldsymbol{v} \right) \right). \tag{9}$$

Lemma 1 *Any optimal solution $\boldsymbol{v}^*$ of the convex problem:*

$$\min_{\|\boldsymbol{v}_s\|_p \leq c, \boldsymbol{v} \in \mathcal{K}} \max_{\boldsymbol{z} \in \mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v} \tag{10}$$

satisfies $\|\boldsymbol{v}_s^*\|_p = c$, i.e., (10) gives an optimal solution of (9), if and only if the optimal value of (9) is negative. Moreover, any optimal $\boldsymbol{v}^*$ of (10) satisfies $\|\boldsymbol{v}_s^*\|_p < c$ if and only if the optimal value of (9) is positive.

Proof We will focus on the first statement because we can prove the second one similarly. We start with proving "if" part. If $\boldsymbol{v}^*$ of (10) satisfies $\|\boldsymbol{v}_s^*\|_p < c$ on contrary, we will get

$$\begin{aligned} \max_{\boldsymbol{z} \in \mathcal{U}} -\boldsymbol{z}^\top \left(\frac{c\boldsymbol{v}^*}{\|\boldsymbol{v}_s^*\|_p} \right) &< \max_{\boldsymbol{z} \in \mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v}^* \\ &= \min_{\|\boldsymbol{v}_s\|_p \leq c, \boldsymbol{v} \in \mathcal{K}} \max_{\boldsymbol{z} \in \mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v} \\ &\leq \min_{\|\boldsymbol{v}_s\|_p = c, \boldsymbol{v} \in \mathcal{K}} \max_{\boldsymbol{z} \in \mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v} < 0, \end{aligned} \tag{11}$$

which contradicts the optimality of $\boldsymbol{v}^*$.

Now we show that the inverse also holds true. If the optimal value of (9) is not negative, then we have

$$\min_{\|\boldsymbol{v}_s\|_p \leq c, \boldsymbol{v} \in \mathcal{K}} \max_{\boldsymbol{z} \in \mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v} = \min_{\|\boldsymbol{v}_s\|_p = c, \boldsymbol{v} \in \mathcal{K}} \max_{\boldsymbol{z} \in \mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v} \geq 0,$$

which means that $\boldsymbol{v}^* = \mathbf{0}$ is an optimal solution of the convex problem (10), but it violates $\|\boldsymbol{v}_s^*\|_p = c$ for $c > 0$. The optimal value of (9) is certainly negative. □

Remark 1 We apply Theorem 1 to the case $V = \mathcal{K}$. Then there are only two possible values (0 or ∞) for the threshold τ if the optimal value (9) is not zero. If (9) has the negative optimal value, the optimal value of (5) goes $-\infty$ and $\tau = \infty$. In this case, $c < \tau$ always holds and solving the convex problem (10) is sufficient for (9).

On the other hand, when (9) has the positive optimal value, an optimal solution of (5) is $\mathbf{0}$ and, therefore, $\tau = 0$. In this case, $c < \tau$ never holds and we have to deal with the nonconvex problem (9). We see that the claims of Lemma 1 are consistent with Theorem 1.

In the following subsections, we relate the positivity (or negativity) of the optimal value of (9) to a geometric condition to $\mathcal{U}$.

3.2.1 $\mathcal{K} = \mathbb{R}^n$

We assume a special cone for $\mathcal{K}$ as $\mathcal{K} = \mathbb{R}^n$ and relate the statements of Lemma 1 to a geometric condition for $\mathcal{U}$.

Theorem 2 *These two statements hold:*

(i) $\mathbf{0} \notin \mathcal{U}$ *if and only if any optimal solution of the convex problem* (10) *is optimal to* (9).
(ii) $\mathbf{0}$ *is in the interior of* $\mathcal{U}$*, i.e.,* $\mathbf{0} \in \mathrm{int}(\mathcal{U})$ *if and only if any optimal solution of the convex problem* (10) *is not optimal to* (9).

Proof By Lemma 1, it is sufficient to show that the optimal value of (9) with $\mathcal{K} = \mathbb{R}^n$ is negative iff $\mathbf{0} \notin \mathcal{U}$ and that it is positive iff $\mathbf{0} \in \mathrm{int}(\mathcal{U})$ for proving this theorem. We can prove them by following the proof of Lemma 1 in [21], but to make our paper self-contained, we roughly sketch the proof.

Suppose that $\mathbf{0} \notin \mathcal{U}$. The simplest separation theorem shows that there is a hyperplane (e.g., $\{z : -z^\top \hat{v} = \bar{c}\}$) that separates convex $\mathcal{U}$ and $\{\mathbf{0}\}$ strictly, i.e.,

$$\exists \hat{v} \neq \mathbf{0} \text{ s.t. } -z^\top \hat{v} < \bar{c} < -\mathbf{0}^\top \hat{v} = 0 \quad \text{for } \forall z \in \mathcal{U},$$

which implies $\max_{z \in \mathcal{U}} -z^\top \hat{v} < 0$. This proves the negativeness of the optimal value of (9) with $\mathcal{K} = \mathbb{R}^n$. When $\mathbf{0} \in \mathrm{int}(\mathcal{U})$, any v $(\neq \mathbf{0})$ obviously achieves $\max_{z \in \mathcal{U}} -z^\top v > 0$, which implies the positiveness of (9). Finally, we think of the case that $\mathbf{0}$ exists in the boundary of $\mathcal{U}$, i.e., $\mathbf{0} \in \mathrm{bd}(\mathcal{U})$. Then there exists a supporting hyperplane to $\mathcal{U}$ at $\mathbf{0}$, that leads to $\min_{v \neq 0} \max_{z \in \mathcal{U}} -z^\top v = 0$. This implies the zero optimal value for (9) with $\mathcal{K} = \mathbb{R}^n$.

In the above, we have proved that the position of $\mathbf{0}$ relative to $\mathcal{U}$ determines the sign of the optimal value of (9). By taking the contrapositive of the above statements, we can ensure that the inverse statements also hold true, i.e., the position of $\mathbf{0}$ relative to $\mathcal{U}$ is determined from the optimal value of (9). □

3.2.2 Convex cone $\mathcal{K}$

We assume a convex cone $\mathcal{K}$ for V. In this case, Theorem 2 is weaken as follows:

Theorem 3 *If* $\mathbf{0} \in \mathrm{int}(\mathcal{U})$ *holds,* (10) *does not give an optimal solution of* (9)*. Moreover, if* (10) *gives an optimal solution of* (9)*,* $\mathbf{0} \notin \mathcal{U}$*.*

Proof The first statement corresponds to "only if" in (ii) and the second statement corresponds to "if" in (i) in Theorem 2 (in the case of general convex cones, the inverse does not necessarily hold). For the inequality:

$$\min_{\|\boldsymbol{v}_s\|_p=c} \max_{z\in\mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v} \leq \min_{\|\boldsymbol{v}_s\|_p=c, \boldsymbol{v}\in\mathcal{K}} \max_{z\in\mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v}, \tag{12}$$

the condition $\mathbf{0} \in \mathrm{int}(\mathcal{U})$ ensures the positivity of the left hand side problem, i.e., (9) with $\mathcal{K} = \mathbb{R}^n$ (see Lemma 1 and Theorem 2). Therefore, the optimal value of the right hand side problem, (9), is also positive, for which Lemma 1 proves that (10) does not give an optimal solution of (9).

Similarly, if (10) gives an optimal solution of (9), the optimal value of the right hand side problem in (12) must be negative. Then the optimal value of (9) with $\mathcal{K} = \mathbb{R}^n$ is also negative, and therefore, we obtain the condition $\mathbf{0} \notin \mathcal{U}$ from Theorem 2. □

4 Mixed integer optimization approach

4.1 Examples of uncertainty sets $\mathcal{U}$

Theorem 3.1 of [2] implies that if μ is coherent, the corresponding $\mathcal{Q}$ to μ leads to $\mu(\tilde{\boldsymbol{z}}^\top \boldsymbol{v}) = \max_{z\in\mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v}$ with the use of $\mathcal{U}$ of (4). Theorem 4 of [11] similarly shows that for a coherent risk measure, there exists $\mathcal{U}$ such that $\mathcal{U} \subseteq \mathrm{conv}(\mathcal{Z})$. Moreover, the theorem mentions that the inverse holds as well, i.e. the risk measure $\mu(\tilde{\boldsymbol{z}}^\top \boldsymbol{v})$ defined by $\max_{z\in\mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v}$ with a convex set $\mathcal{U} \subseteq \mathrm{conv}(\mathcal{Z})$ is coherent. In other words, the theorem provides a way of making a risk measure corresponding to $\mathcal{U}$ coherent by replacing $\mathcal{U}$ by $\mathcal{U} \cap \mathrm{conv}(\mathcal{Z})$. We will show a new coherent risk measure by preparing a set $\mathcal{U}$ satisfying $\mathcal{U} \subseteq \mathrm{conv}(\mathcal{Z})$ with the use of a Bregman divergence in Example 4 and investigate the performance of portfolio models based on minimizing the new coherent risk measure in numerical results.

The following examples show examples of $\mathcal{U}$ derived from probability sets $\mathcal{Q}$ and confirm theorems of [2,11] for such $\mathcal{U}$. Note that when $\mathcal{Q} \subseteq \Delta^m$, $\mathcal{U}$ constructed by (4) satisfies the condition of coherent risk measures, $\mathcal{U} \subseteq \mathrm{conv}(\mathcal{Z})$.

Example 1 (Scenario-based Set) We consider the coherent risk measure generated by

$$\mathcal{Q} = \mathrm{conv}\left(\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k\}\right),$$

where $\boldsymbol{q}_i \in \Delta^m$. Then the corresponding uncertainty set is $\mathcal{U} = \mathrm{conv}(\{\boldsymbol{Z}\boldsymbol{q}_1, \ldots, \boldsymbol{Z}\boldsymbol{q}_k\})$.

Example 2 (Conditional Value-at-Risk (CVaR)) The coherent risk measure known as CVaR has the generating family $\mathcal{Q} = \{\boldsymbol{q} : \boldsymbol{q} \in \Delta^m, \boldsymbol{q} \leq \frac{\hat{\boldsymbol{p}}}{\nu}\}$ for the given $\hat{\boldsymbol{p}}$. Here let ν be a parameter that takes the value in $(0, 1]$. $\mathcal{Q}$ can be regarded as the uncertainty distribution set for $\hat{\boldsymbol{p}}$ because $\mathcal{Q}$ is the set of probabilities with center at $\hat{\boldsymbol{p}}$. From (4), we get

$$\mathcal{U} = \mathrm{conv}\left(\{\boldsymbol{Z}\boldsymbol{q} : \boldsymbol{q} \in \mathcal{Q}\}\right) = \left\{\sum_i q_i \boldsymbol{z}_i : \boldsymbol{q} \in \Delta^m, \boldsymbol{q} \leq \frac{\hat{\boldsymbol{p}}}{\nu}\right\}. \tag{13}$$

Note that the size of $\mathcal{U}$ is monotonically decreasing with respect to ν (see (a) in Fig. 1). By taking dual to the inner problem in (3), we have the following problem equivalent to (3):

$$\begin{aligned} \min_{\boldsymbol{v}\in V,\alpha,\boldsymbol{\xi}} \quad & \alpha + \frac{1}{\nu}\sum_{i=1}^{m} \hat{p}_i \xi_i \\ \text{s.t.} \quad & \xi_i + \boldsymbol{z}_i^\top \boldsymbol{v} + \alpha \geq 0,\ i = 1, \ldots, m, \\ & \boldsymbol{\xi} \geq 0,\ \|\boldsymbol{v}_s\|_p = c, \end{aligned} \tag{14}$$

which minimizes the CVaR defined in the discrete distribution $\{-\boldsymbol{z}_1^\top \boldsymbol{v}, \ldots, -\boldsymbol{z}_m^\top \boldsymbol{v}\}$.

When $\nu = 1$, we have $\mathcal{Q} = \{\hat{\boldsymbol{p}}\}$. Then $\mathcal{U} = \{\sum_{i=1}^m \hat{p}_i z_i\}$, and $\mu(\tilde{\boldsymbol{z}}^\top \boldsymbol{v})$ is equivalent to the expectation of $-\boldsymbol{z}_i^\top \boldsymbol{v}$, that is, $\sum_{i=1}^m \hat{p}_i(-\boldsymbol{z}_i^\top \boldsymbol{v})$.

When ν is sufficiently close to 0, we have $\mathcal{Q} = \{\boldsymbol{q} \in \Delta^m\} = \mathrm{conv}(\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m\})$, where $\boldsymbol{e}_i$ is the ith unit coordinate vector. Then $\mathcal{U} = \mathrm{conv}(\mathcal{Z})$ and $\mu(\tilde{\boldsymbol{z}}^\top \boldsymbol{v})$ is equivalent to the worst-case scenario of $-\boldsymbol{z}_i^\top \boldsymbol{v}$, that is, $\max_{i=1,\ldots,m}(-\boldsymbol{z}_i^\top \boldsymbol{v})$.

Example 3 (One-sided Moments) Consider the coherent risk measure:

$$\begin{aligned} \mu\left(\tilde{\boldsymbol{z}}^\top \boldsymbol{v}\right) &= -\mathbb{E}_{\hat{\boldsymbol{p}}}\left(\tilde{\boldsymbol{z}}^\top \boldsymbol{v}\right) + \lambda\left[\mathbb{E}_{\hat{\boldsymbol{p}}}\left(\left[-\tilde{\boldsymbol{z}}^\top \boldsymbol{v} + \mathbb{E}_{\hat{\boldsymbol{p}}}\left(\tilde{\boldsymbol{z}}^\top \boldsymbol{v}\right)\right]^+\right)^r\right]^{1/r}, \\ &\text{where } [a]^+ = \max\{a, 0\}, \end{aligned}$$

represented by the families of measures

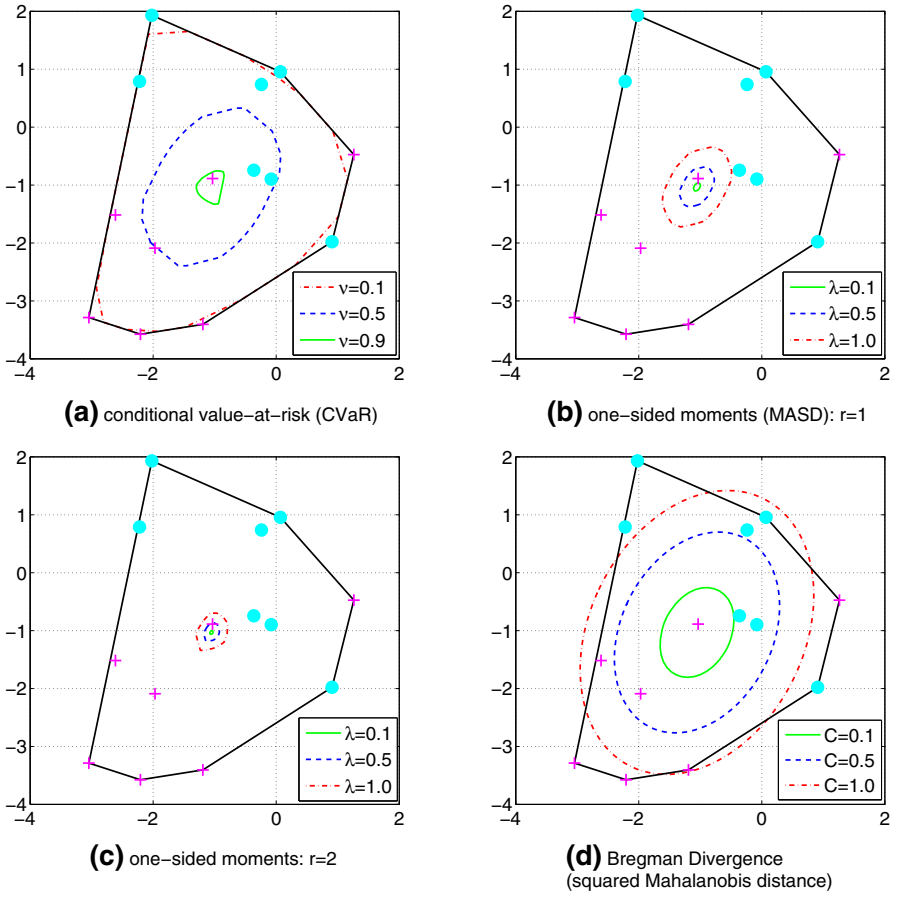
$$\mathcal{Q} = \{\boldsymbol{q} : q_i = \hat{p}_i\left(1 + \lambda\left(u_i - \boldsymbol{u}^\top \hat{\boldsymbol{p}}\right)\right),\ \boldsymbol{0} \leq \boldsymbol{u},\ \|\boldsymbol{u}\|_{\tilde{r}} \leq 1\} \tag{15}$$

for $r \geq 1$, $0 \leq \lambda \leq 1$ and $\tilde{r} = r/(r-1)$. When $r = 1$ (i.e., $\tilde{r} = \infty$), the risk measure is known as mean absolute semi-deviation (MASD). $\mathcal{Q}$ coincides with $\hat{\boldsymbol{p}}$ when $\lambda = 0$ and it is monotonically increasing with respect to λ (see (b) and (c) in Fig. 1). It is known that when $\lambda \in [0, 1]$, the risk measure is coherent (see [8]). The corresponding uncertainty set $\mathcal{U}$ is described as

$$\mathcal{U} = \left\{z : z = \bar{z} + \lambda\left(\boldsymbol{Z}\boldsymbol{P}\boldsymbol{u} - \left(\boldsymbol{u}^\top \hat{\boldsymbol{p}}\right)\bar{z}\right),\ \boldsymbol{0} \leq \boldsymbol{u},\ \|\boldsymbol{u}\|_{\tilde{r}} \leq 1\right\},$$

where $\bar{\boldsymbol{z}}$ is the mean for $\mathcal{Z}$ and $\boldsymbol{P}$ is the diagonal matrix with diagonal entries $\hat{\boldsymbol{p}}$. λ such as $\lambda > 1$ may leads to an unacceptable set of probabilities $\mathcal{Q}$, since some of $\boldsymbol{q} \in \mathcal{Q}$ include negative valued components. Therefore, to make $\mathcal{U}$ larger by increasing λ to more than 1, it may be reasonable to add $\boldsymbol{q} \geq \boldsymbol{0}$ to $\mathcal{Q}$ of (15), though the modified $\mathcal{Q}$ changes the type of concerned risk measures.

Example 4 (Bregman divergence) Here we show a new coherent risk measure by preparing a set $\mathcal{U}$ satisfying $\mathcal{U} \subseteq \mathrm{conv}(\mathcal{Z})$ with the use of a Bregman divergence. Let $F : \Omega \to \mathbb{R}$ be a continuously-differentiable real-valued and strictly convex function defined on a closed convex set Ω. The Bregman distance associated with F for points,



(a) conditional value–at–risk (CVaR)

(b) one–sided moments (MASD): r=1

(c) one–sided moments: r=2

(d) Bregman Divergence (squared Mahalanobis distance)

Fig. 1 Uncertainty sets $\mathcal{U}$ for several risk measures. The plots *filled circle* and *plus* indicate data points with the label "+1" and "−1", respectively, for binary classification (see Sect. 5.2)

$\hat{\boldsymbol{p}}$ and $\boldsymbol{q} \in \Omega$, is the difference between the value of F at point $\boldsymbol{q}$ and the value of the first-order Taylor expansion of F around point $\hat{\boldsymbol{p}}$ evaluated at point $\boldsymbol{q}$:

$$D_F^p(\boldsymbol{q}, \hat{\boldsymbol{p}}) = F(\boldsymbol{q}) - F(\hat{\boldsymbol{p}}) - \nabla F(\hat{\boldsymbol{p}})^\top (\boldsymbol{q} - \hat{\boldsymbol{p}}).$$

Using special cases of F, we have

$$\begin{array}{ll} \text{Squared Euclidean:} & D_F^p(\boldsymbol{q}, \hat{\boldsymbol{p}}) = \|\boldsymbol{q} - \hat{\boldsymbol{p}}\|^2 \text{ by } F(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 \\ \text{Squared Mahalanobis:} & D_F^p(\boldsymbol{q}, \hat{\boldsymbol{p}}) = \frac{1}{2}(\boldsymbol{q} - \hat{\boldsymbol{p}})^\top \boldsymbol{M}(\boldsymbol{q} - \hat{\boldsymbol{p}}) \text{ by } F(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} \\ \text{Kullback-Leibler:} & D_F^p(\boldsymbol{q}, \hat{\boldsymbol{p}}) = \sum_{i=1}^m q_i \left(\ln \frac{q_i}{\hat{p}_i}\right) \text{ by } F(\boldsymbol{x}) = \sum x_i \ln x_i. \end{array}$$

Here, let $\bar{\boldsymbol{z}}$ and $\boldsymbol{\Sigma}_z$ be the mean and the covariance matrix, respectively, for $\mathcal{Z}$. We assume that $\boldsymbol{\Sigma}_z$ is invertible and define $\boldsymbol{M}$ of the squared Mahalanobis distance by $\boldsymbol{M} = \boldsymbol{Z}^\top \boldsymbol{\Sigma}_z^{-1} \boldsymbol{Z}$. Here we define

$$\mathcal{Q} = \left\{ \boldsymbol{q} \in \Delta^m : D_F^p\left(\boldsymbol{q}, \hat{\boldsymbol{p}}\right) \leq C \right\} = \left\{ \boldsymbol{q} \in \Delta^m : \frac{1}{2} \left(\boldsymbol{Z}\boldsymbol{q} - \bar{\boldsymbol{z}}\right)^\top \boldsymbol{\Sigma}_z^{-1} \left(\boldsymbol{Z}\boldsymbol{q} - \bar{\boldsymbol{z}}\right) \leq C \right\}$$

with using some $C > 0$. Then the corresponding $\mathcal{U}$ is described as follows:

$$\begin{aligned} \mathcal{U} &= \left\{ \boldsymbol{Z}\boldsymbol{q} : \boldsymbol{q} \in \Delta^m, \frac{1}{2} \left(\boldsymbol{Z}\boldsymbol{q} - \bar{\boldsymbol{z}}\right)^\top \boldsymbol{\Sigma}_z^{-1} \left(\boldsymbol{Z}\boldsymbol{q} - \bar{\boldsymbol{z}}\right) \leq C \right\} \\ &= \mathrm{conv}(\mathcal{Z}) \cap \left\{ \boldsymbol{z} : \frac{1}{2} \left(\boldsymbol{z} - \bar{\boldsymbol{z}}\right)^\top \boldsymbol{\Sigma}_z^{-1} \left(\boldsymbol{z} - \bar{\boldsymbol{z}}\right) \leq C \right\}. \end{aligned} \tag{16}$$

$\mathcal{U}$ is the intersection of the convex hull of $\mathcal{Z}$ and the ellipsoid with center $\bar{\boldsymbol{z}}$ and shape described by $\boldsymbol{\Sigma}_z^{-1}$ (see (d) in Fig. 1).

Consider the uncertainty set which consists of a quadratic constraint in (16):

$$\widehat{\mathcal{U}} = \left\{ \boldsymbol{z} : \frac{1}{2} \left(\boldsymbol{z} - \bar{\boldsymbol{z}}\right)^\top \boldsymbol{\Sigma}_z^{-1} \left(\boldsymbol{z} - \bar{\boldsymbol{z}}\right) \leq C \right\}. \tag{17}$$

The robust optimization (3) with the uncertainty set $\widehat{\mathcal{U}}$ reduces to

$$\min_{\|\boldsymbol{v}_s\|_p = c, \boldsymbol{v} \in V} -\bar{\boldsymbol{z}}^\top \boldsymbol{v} + \alpha \sqrt{\boldsymbol{v}^\top \boldsymbol{\Sigma}_z \boldsymbol{v}},$$

where $\alpha = \sqrt{2C}$. It is known as a classical mean-standard deviation portfolio allocation problem because $-\bar{\boldsymbol{z}}^\top \boldsymbol{v}$ and $\sqrt{\boldsymbol{v}^\top \boldsymbol{\Sigma}_z \boldsymbol{v}}$ are the expected value and the standard deviation of the random portfolio return $-\tilde{\boldsymbol{z}}^\top \boldsymbol{v}$. The mean-standard deviation risk measure is not a coherent risk measure, but $\mathcal{U} = \mathrm{conv}(\mathcal{Z}) \cap \widehat{\mathcal{U}}$ of (16) leads to a coherent risk measure because $\mathcal{U} \subseteq \mathrm{conv}(\mathcal{Z})$ (see [11]).

4.2 Integer optimization formulation for ℓ_1/ℓ_∞-norm problem

We showed several examples of $\mathcal{Q}$ for coherent risk minimization problem (2). Most of them can be described in a linear representation or conic representation:

$$\mathcal{Q}_p = \left\{ \boldsymbol{q} \in \Delta^m : \boldsymbol{A}\boldsymbol{q} \leq \boldsymbol{b} \right\} \text{ or } \mathcal{Q}_c = \left\{ \boldsymbol{q} \in \Delta^m : \|\boldsymbol{B}\boldsymbol{q}\|_{p'} \leq 1 \right\} \tag{18}$$

except for Kullback-Leibler distance in Example 4. We focus on $\mathcal{Q}_p$ and $\mathcal{Q}_c$ and deal with the resulting coherent risk minimization problems (2). When $\mathcal{U}$ is defined for $\mathcal{Q} = \mathcal{Q}_p$ as

$$\mathcal{U} = \left\{ \boldsymbol{Z}\boldsymbol{q} : \boldsymbol{q} \in \Delta^m, \boldsymbol{A}\boldsymbol{q} \leq \boldsymbol{b} \right\},$$

(3) reduces to

$$\begin{aligned} \min_{\boldsymbol{v} \in V, \alpha, \boldsymbol{\xi}} \quad & \alpha + \boldsymbol{b}^\top \boldsymbol{\xi} \\ \text{s.t.} \quad & \boldsymbol{Z}^\top \boldsymbol{v} + \alpha \boldsymbol{e} + \boldsymbol{A}^\top \boldsymbol{\xi} \geq 0, \ \boldsymbol{\xi} \geq \boldsymbol{0}, \\ & \|\boldsymbol{v}_s\|_p = c. \end{aligned} \tag{19}$$

When $\mathcal{U}$ is defined for $\mathcal{Q} = \mathcal{Q}_c$ as

$$\mathcal{U} = \{\boldsymbol{Zq} : \boldsymbol{q} \in \Delta^m, \|\boldsymbol{Bq}\|_{p'} \leq 1\},$$

(3) reduces to

$$\begin{array}{ll} \min\limits_{\boldsymbol{v} \in V, \alpha, \boldsymbol{\xi}} & \alpha + \|\boldsymbol{\xi}\|_{q'} \\ \text{s.t.} & \boldsymbol{Z}^\top \boldsymbol{v} + \alpha \boldsymbol{e} - \boldsymbol{B}^\top \boldsymbol{\xi} \geq 0, \\ & \|\boldsymbol{v}_s\|_p = c. \end{array} \tag{20}$$

$\|\cdot\|_{q'}$ indicates the dual norm of the $\ell_{p'}$ norm ($\frac{1}{p'} + \frac{1}{q'} = 1$).

Problems (19) and (20) seem difficult to solve due to the nonconvex constraint: $\|\boldsymbol{v}_s\|_p = c$. However, Theorem 1 shows a condition for reducing the nonconvex problem (3) to a convex problem for any ℓ_p norm. It implies that if c is smaller than some threshold (that is defined by solving a convex optimization problem), we can obtain an optimal solution of (3) by solving a convex relaxation problem of (3) whose norm constraint is $\|\boldsymbol{v}_s\|_p \leq c$. Therefore, before solving (19) or (20), it might be good to check whether c of the problem satisfies the condition.

If c is larger than the threshold, we need to apply nonconvex optimization techniques to solve (19) or (20). For the ℓ_2 norm, it might be difficult to reformulate the problem into a solvable problem. Here we will use the ℓ_1 or ℓ_∞ norm for $\|\boldsymbol{v}_s\|_p = c$ and formulate (19) and (20) as solvable problems, mixed integer optimization (MIO) problems.

For that purpose, we introduce new binary variables $\boldsymbol{u}$ and new nonnegative variables $\boldsymbol{v}^+, \boldsymbol{v}^-$ which replace $\boldsymbol{v}$. On condition that at least either of v_i^+ and v_i^- must be zero for every element $i \in \mathcal{S}$, we identify $\boldsymbol{v}$ and $\boldsymbol{v}^+ - \boldsymbol{v}^-$ and describe $\|\boldsymbol{v}_s\|_1$ by $\sum_{i \in \mathcal{S}} (v_i^+ + v_i^-)$. Then we reformulate (19) with the ℓ_1 norm as follows:

$$\begin{array}{ll} \min\limits_{\boldsymbol{v}^+, \boldsymbol{v}^-, \alpha, \boldsymbol{\xi}, \boldsymbol{u}} & \alpha + \boldsymbol{b}^\top \boldsymbol{\xi} \\ \text{s.t.} & \boldsymbol{Z}^\top (\boldsymbol{v}^+ - \boldsymbol{v}^-) + \alpha \boldsymbol{e} + \boldsymbol{A}^\top \boldsymbol{\xi} \geq \boldsymbol{0}, \quad \boldsymbol{\xi} \geq \boldsymbol{0}, \\ & (\boldsymbol{v}^+ - \boldsymbol{v}^-) \in V, \\ & \sum\limits_{i \in \mathcal{S}} (v_i^+ + v_i^-) = c, \\ & 0 \leq v_i^+ \leq c u_i, \quad 0 \leq v_i^- \leq c(1 - u_i), \quad u_i \in \{0, 1\}, \ i \in \mathcal{S}. \end{array} \tag{21}$$

If (19) with the ℓ_1 norm satisfies the transferability condition for c, (19) equivalently reduces to the convex relaxation problem of (21) where the constraints involving the integer variables $u_i \in \{0, 1\}$ are all removed from (21) and the constraint $\|\boldsymbol{v}_s\|_1 = c$ is simply replaced by

$$\sum_{i \in \mathcal{S}} \left(v_i^+ + v_i^-\right) \leq c.$$

When $\mathcal{U}$ of (3) is conic representable, the robust optimization problem (20) with the ℓ_1 norm results in a conic integer optimization problem:

$$\begin{aligned}\min_{\boldsymbol{v}^+,\boldsymbol{v}^-,\alpha,\boldsymbol{\xi},\boldsymbol{u}} \quad & \alpha + \|\boldsymbol{\xi}\|_{q'} \\ \text{s.t.} \quad & \boldsymbol{Z}^\top(\boldsymbol{v}^+ - \boldsymbol{v}^-) + \alpha \boldsymbol{e} - \boldsymbol{B}^\top \boldsymbol{\xi} \geq \boldsymbol{0}, \\ & (\boldsymbol{v}^+ - \boldsymbol{v}^-) \in V, \\ & \sum_{i \in \mathcal{S}} (v_i^+ + v_i^-) = c, \\ & 0 \leq v_i^+ \leq c u_i, \ \ 0 \leq v_i^- \leq c(1 - u_i), \ \ u_i \in \{0, 1\}, \ i \in \mathcal{S}.\end{aligned}$$

We can also transform (19) and (20) into MIO problems in the case of the ℓ_∞-norm constraint $\|\boldsymbol{v}_s\|_\infty = c$. As well as the ℓ_1-norm case, we identify $\boldsymbol{v}$ and $\boldsymbol{v}^+ - \boldsymbol{v}^-$ on condition that at least either of v_i^+ and v_i^- must be zero for every element $i \in \mathcal{S}$. The constraint $\|\boldsymbol{v}_s\|_\infty = \max_{i \in \mathcal{S}} |v_i| = c$ requires that $v_i^+ + v_i^- = c$ holds at least one of $i \in \mathcal{S}$. Therefore, it can be described by $v_i^+ + v_i^- \geq c r_i$ with the use of additional binary variables $r_i \in \{0, 1\}$ for $i \in \mathcal{S}$ and $\sum_{i \in \mathcal{S}} r_i \geq 1$. As a result, (19) with the ℓ_∞ norm can be reformulated as

$$\begin{aligned}\min_{\boldsymbol{v}^+,\boldsymbol{v}^-,\alpha,\boldsymbol{\xi},\boldsymbol{u},\boldsymbol{r}} \quad & \alpha + \boldsymbol{b}^\top \boldsymbol{\xi} \\ \text{s.t.} \quad & \boldsymbol{Z}^\top(\boldsymbol{v}^+ - \boldsymbol{v}^-) + \alpha \boldsymbol{e} + \boldsymbol{A}^\top \boldsymbol{\xi} \geq \boldsymbol{0}, \ \boldsymbol{\xi} \geq \boldsymbol{0}, \\ & (\boldsymbol{v}^+ - \boldsymbol{v}^-) \in V, \\ & v_i^+ + v_i^- \geq c r_i, \ r_i \in \{0, 1\}, \ i \in \mathcal{S}, \\ & \sum_{i \in \mathcal{S}} r_i \geq 1, \\ & 0 \leq v_i^+ \leq c u_i, \ 0 \leq v_i^- \leq c(1 - u_i), \ u_i \in \{0, 1\}, \ i \in \mathcal{S}.\end{aligned} \tag{22}$$

The ℓ_∞-norm version of (20) is similarly formulated as an MIO problem.

5 Applications of norm-constrained coherent risk minimization

5.1 Portfolio optimization

Portfolio optimization is a problem of determining a normalized weight vector $\boldsymbol{v}$ to make the portfolio better than any other according to some criterion. As such a criterion, we can use a coherent risk measure. We assume that $\tilde{\boldsymbol{z}}$ is a random portfolio return vector and find the best allocation $\boldsymbol{v}$ by minimizing a risk measure.

5.1.1 Portfolio optimization of hedge funds

We consider the following portfolio selection problem:

$$\min_{\|\boldsymbol{v}\|_1 = c, \boldsymbol{v} \in V} \mu\left(\tilde{\boldsymbol{z}}^\top \boldsymbol{v}\right), \tag{23}$$

where $V = \{\boldsymbol{v} : \boldsymbol{e}^\top \boldsymbol{v} = 1\}$ and c is a positive value larger than 1. The use of borrowed capital to increase the potential return of an investment is called leverage in finance. Leverage can be regarded as the financial risk relative to the invested capital. The

constraint $\|\boldsymbol{v}\|_1 = c$ together with $\boldsymbol{v} \in V$ controls the leverage ratio, described by

$$\frac{\sum_{i\in I} |v_i|}{\sum_{i\in I} v_i} = \sum_{i\in I_+} v_i - \sum_{j\in I_-} v_j = \|\boldsymbol{v}\|_1,$$

where I indicates the index set of all assets, I_+ indicates the index set of assets in long positions, i.e., $v_i \geq 0$ for $i \in I_+$, and I_- indicates the index set of assets in short positions, i.e., $v_j < 0$ for $j \in I_-$. The first equality in the above equations is due to the constraint $\boldsymbol{e}^\top \boldsymbol{v} = 1$. Therefore, the constraint $\|\boldsymbol{v}\|_1 = c$ of (23) requires that the leverage ratio equals to c. The constraint $\|\boldsymbol{v}\|_1 = c$ together with $\boldsymbol{e}^\top \boldsymbol{v} = 1$ can be rewritten as

$$\sum_{i\in I_+} v_i = \frac{c+1}{2} \quad \text{and} \quad \sum_{i\in I_-} |v_i| = \frac{c-1}{2}.$$

The above equations limit the total long holdings and total short holdings. Therefore, the nonconvex leverage constraint has a role of controlling total long and short holdings.

As a common source of nonconvexity in practical portfolio optimization problems, [18] has referred to leverage requirement in addition to threshold constraints on the holdings or trades. However, at the same time, it recommended that such constraints be left out of analyses because there is no theory to support the required optimality conditions. If we formulate the nonconvex problem as an MIO formulation (e.g., (21)) and the resulting MIO has the proper problem size for a highly optimized state-of-the-art MIO solver such as CPLEX, we can obtain a global optimal solution for the nonconvex portfolio allocation problem including leverage requirement. Especially when the nonconvex problem (3) satisfies a criterion of the transferability to the convex relaxation problem, it is sufficient to solve the convex problem (7).

The convex relaxation problem (7) equals to the norm-constraint portfolio optimization model:

$$\min_{\boldsymbol{v}\in V} \ \mu\left(\tilde{\boldsymbol{z}}^\top \boldsymbol{v}\right) \quad \text{s.t. } \|\boldsymbol{v}\|_p \leq c. \tag{24}$$

The model (24) has recently been studied by various researchers. DeMiguel et al. [7] used variance as the risk measure in the norm-constrained portfolio optimizations with various types of norms (note that variance is not coherent), while Gotoh and Takeda [9] used CVaR risk measure. Brodie et al. [5] incorporated an ℓ_1-norm penalty on the portfolio decision vector into the traditional Markowitz portfolio optimization model in order to encourage sparse portfolios.

The constraint $\|\boldsymbol{v}\|_p \leq c$ is meaningful when the inequality constraint becomes active at the optimality. In that sense, we can say that the proposed model (1) embraces existing norm-constraint portfolio optimization models minimizing coherent risk measures.

To make the portfolio optimization model (23) more realistic, we will add constraints that set bounds on the holdings within each market sector to V. There are major sectors of the stock market. For example, the constituents of TOPIX (Tokyo

Stock Price Index) are divided into the 17 categories[1] according to Securities Identification Code Committee (SICC). Since each sector contains several related industries, it is reasonable to limit the holdings or trades on each sector $G_1, \ldots, G_K$. For that purpose, we add the following constraints to V:

$$-\gamma \leq \sum_{i \in G_k} v_i \leq \gamma, \quad \forall k = 1, \ldots, K,$$

for a positive parameter $\gamma > 0$.

5.1.2 Long-only portfolio optimization

We can describe a long-only portfolio optimization model which minimizes a coherent risk measure as (23), where $V = \mathbb{R}^n_+ := \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{v} \geq \boldsymbol{0}\}$ and $c = 1$. This problem is essentially a convex problem because $\|\boldsymbol{v}\|_1 = 1$ can be replaced by $\boldsymbol{e}^\top \boldsymbol{v} = 1$ under the condition that $\boldsymbol{v} \geq \boldsymbol{0}$ in V. However, to make the long-only portfolio optimization model fit to the formulation (3), we have described the model by (23).

5.2 Binary classification method in machine learning

This section shows binary classification models formulated by minimizing coherent risk measures and introduces some results of [21], which formulated a popular soft-margin classification model known as ν-SVM [17] and its extended model [12] by (3) with using $\mathcal{U}$ of (13).

We will introduce additional notations before getting to the main point. Let $\mathcal{X} \subset \mathbb{R}^{n-1}$ be the input domain and $\{+1, -1\}$ be the set of the binary labels. Suppose that we have samples,

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m) \in \mathcal{X} \times \{+1, -1\}.$$

Let M_+ be the index set of $\boldsymbol{x}_i$, $i = 1, \ldots, m$, with the label $+1$ and M_- be the index set of $\boldsymbol{x}_i$, $i = 1, \ldots, m$, with the label -1. We compute $(\boldsymbol{w}, b)$ for a decision function $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ using these samples and use $h(\boldsymbol{x})$ to predict the label for a new input point $\hat{\boldsymbol{x}} \in \mathcal{X}$. If $h(\hat{\boldsymbol{x}})$ is positive (resp. negative), the label of $\hat{\boldsymbol{x}}$ is predicted to be $+1$ (resp. -1). Here we focus on linear learning models using linear functions $h(\boldsymbol{x})$, but the discussions in this section can be directly applicable to non-linear kernel models [16] using nonlinear maps $\phi(\boldsymbol{x})$ from original space to high dimensional space.

[1] Those indices are "Foods", "Energy Resources", "Construction & Materials", "Raw Materials & Chemicals", "Pharmaceutical," "Automobiles & Transportation Equipment", "Steel & Nonferrous Metals", "Machinery", "Electric Appliances & Precision Instruments", "IT & Services, Others", "Electric Power & Gas", "Transportation & Logistics", "Commercial & Wholesale Trade", "Retail Trade", "Banks", "Financials (Ex Banks)" and "Real Estate".

Let the support $\mathcal{Z}$ of $\tilde{\boldsymbol{z}}$ consist of $\boldsymbol{z}_i = \begin{pmatrix} y_i \boldsymbol{x}_i \\ y_i \end{pmatrix}$, $i = 1, \ldots, m$, and let the variable vector $\boldsymbol{v}$ consist of two elements $\boldsymbol{w}$ and b as $\boldsymbol{v} = \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix}$. Now we consider machine learning models which minimize several kinds of coherent risk measures by assuming a family of distributions $\mathcal{Q}$. Suppose that $V = \mathbb{R}^n$ and $\mathcal{S} = \{1, \ldots, n-1\}$ to define $\boldsymbol{v}_s$, i.e., $\boldsymbol{v}_s = \boldsymbol{w}$. We also assume $p = 2$ and $c = 1$ in order to relate (3) to existing machine learning models. Then the problem (3) with $\mathcal{U}$ of (4) results in

$$\min_{\|\boldsymbol{w}\|_2=1, b} \max_{\boldsymbol{q}\in\mathcal{Q}} -\left(\sum_{i\in M_+} q_i \boldsymbol{x}_i - \sum_{j\in M_-} q_j \boldsymbol{x}_j\right)^\top \boldsymbol{w} - \left(\sum_{i\in M_+} q_i - \sum_{j\in M_-} q_j\right) b,$$

and we further transform it into

$$\min_{\|\boldsymbol{w}\|_2=1} \max_{\boldsymbol{q}} -\left(\sum_{i\in M_+} q_i \boldsymbol{x}_i - \sum_{j\in M_-} q_j \boldsymbol{x}_j\right)^\top \boldsymbol{w} \quad \text{s.t. } \boldsymbol{q} \in \mathcal{Q}, \sum_{i\in M_+} q_i = \sum_{j\in M_-} q_j \tag{25}$$

by using the optimality condition for b. Section 4.1.4 in [3] discusses Fisher's linear discriminant (FLD) that minimizes $-(\boldsymbol{x}_+ - \boldsymbol{x}_-)^\top \boldsymbol{w}$ in terms of $\boldsymbol{w} \in \{\boldsymbol{w} : \|\boldsymbol{w}\|_2 = 1\}$ for sample means $\boldsymbol{x}_+, \boldsymbol{x}_-$ of each class so that two classes are well separated. (25) can be regarded as a robust variant of the FLD under distribution uncertainty; uncertain sample means $\boldsymbol{x}_+, \boldsymbol{x}_-$ are described as $\sum_{i\in M_+} q_i \boldsymbol{x}_i$ and $\sum_{j\in M_-} q_j \boldsymbol{x}_j$, respectively. In (3), $\mathcal{U}$ defines the area where uncertain $\boldsymbol{x}_+ - \boldsymbol{x}_-$ can move.

Figure 1 shows the set $\mathcal{U}$ of (4) as

$$\mathcal{U} = \left\{ \sum_{i\in M_+} q_i \boldsymbol{x}_i - \sum_{j\in M_-} q_j \boldsymbol{x}_j : \boldsymbol{q} \in \mathcal{Q}, \sum_{i\in M_+} q_i = \sum_{j\in M_-} q_j \right\}$$

by changing the size parameter of $\mathcal{Q}$ in addition to plotting $\boldsymbol{x}_i$, $i \in M_+$, by "•" and $\boldsymbol{x}_j$, $j \in M_-$, by "+". The uncertain sample means, $\sum_{i\in M_+} q_i \boldsymbol{x}_i$ and $\sum_{j\in M_-} q_j \boldsymbol{x}_j$, can coincide with some $\boldsymbol{q} \in \mathcal{Q}$ when $\mathbf{0} \in \mathcal{U}$. Remind that (3) is essentially nonconvex when $\mathbf{0} \in \mathrm{int}(\mathcal{U})$, while (3) reduces to the convex problem (7) when $\mathbf{0} \notin \mathcal{U}$.

Scenario-based Set: As in Example 1, we consider the coherent risk measure generated by

$$\mathcal{Q} = \mathrm{conv}\,(\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m\}) = \Delta^m.$$

Then from (3), we get

$$\min_{\|\boldsymbol{w}\|_2=1, b} \max_{\boldsymbol{z}\in\mathcal{U}} -\boldsymbol{z}^\top \boldsymbol{v}, \tag{26}$$

where $\mathcal{U} = \text{conv}(\mathcal{Z})$. The problem is rewritten as

$$\max_{\boldsymbol{w},b} \min_{i=1,\ldots,m} \frac{y_i\left(\boldsymbol{x}_i^\top \boldsymbol{w} + b\right)}{\|\boldsymbol{w}\|_2},$$

which is a formulation of hard margin SVM [6].

Conditional Value-at-Risk: We consider the CVaR, which is generated by $\mathcal{Q} = \{\boldsymbol{q} : \boldsymbol{q} \in \Delta^m, \boldsymbol{q} \le \frac{1}{\nu m}\boldsymbol{e}\}$ with parameter $\nu \in (0, 1]$ as in Example 2 ($\hat{\boldsymbol{p}}$ is set to $\frac{1}{m}\boldsymbol{e}$ and the ℓ_2 norm is adopted). The problem (14) can be written as

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi},\alpha} \quad & \nu\alpha + \frac{1}{m}\sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b) \ge -\alpha - \xi_i,\ i = 1, \ldots, m, \\ & \boldsymbol{\xi} \ge \boldsymbol{0},\ \|\boldsymbol{w}\|_2 = 1, \end{aligned} \tag{27}$$

that is a formulation of Eν-SVM [12] (see [21]). The convex relaxation problem of (27), which has $\|\boldsymbol{w}\|_2 \le 1$ instead of $\|\boldsymbol{w}\|_2 = 1$, is corresponding to one of popular soft-margin classification models, ν-SVM, proposed by [17].

Theorem 2 shows that if $\boldsymbol{0} \notin \mathcal{U}$, it is enough to solve the convex optimization problem (10); therefore, one would solve ν-SVM rather than Eν-SVM when $\mathcal{U}$ of (13), defined with $\mathcal{Q} = \{\boldsymbol{q} : \boldsymbol{q} \in \Delta^m, \boldsymbol{q} \le \frac{1}{\nu m}\boldsymbol{e}\}$, satisfies $\boldsymbol{0} \notin \mathcal{U}$ for a given ν. When we find that $\boldsymbol{0} \in \mathcal{U}$, we need to deal with (9) (or (27) in this application) as a nonconvex problem. How can we check whether $\boldsymbol{0} \in \mathcal{U}$ or not? The following shows an example how to check it using $\mathcal{U}$ of the CVaR defined as (13) in Example 2.

Remark 2 The uncertainty set $\mathcal{U}$ of (13) includes a parameter $\nu \in (0, 1]$, which controls the size of the set. As ν becomes larger, $\mathcal{U}$ becomes smaller. That is, as ν is sufficiently larger, $\mathcal{U}$ shrinks to $\hat{\boldsymbol{p}} \ne 0$, which implies that $\boldsymbol{0}$ is not included in $\mathcal{U}$ with sufficiently large ν (see Fig. 1a). Therefore, when conv($\mathcal{Z}$) includes $\boldsymbol{0}$, we can find a threshold $\hat{\nu}$ where $\mathcal{U}$ includes $\boldsymbol{0}$ in its boundary by solving the following linear optimization problem:

$$\begin{aligned} \min_{\phi,\boldsymbol{q}} \quad & \phi \\ \text{s.t.} \quad & \sum_i q_i z_i = \boldsymbol{0},\ \ \boldsymbol{q} \in \Delta^m,\ \ \boldsymbol{q} \le \phi\hat{\boldsymbol{p}}. \end{aligned}$$

For the application to binary classification, the problem is rewritten as

$$\begin{aligned} \min_{\phi,\boldsymbol{q}} \quad & \phi \\ \text{s.t.} \quad & \sum_{i\in M_+} q_i \boldsymbol{x}_i = \sum_{j\in M_-} q_j \boldsymbol{x}_j,\ \ \sum_{i\in M_+} q_i = \sum_{j\in M_-} q_j, \\ & \boldsymbol{e}^\top \boldsymbol{q} = 1,\ \ \boldsymbol{0} \le \boldsymbol{q} \le \tfrac{\phi}{m}\boldsymbol{e} \end{aligned} \tag{28}$$

using the definition of z_i. We can compute $\hat{\nu}$ by $\hat{\nu} = \frac{1}{\phi^*}$ from the optimal solution ϕ^* of the above problem. If conv($\mathcal{Z}$) do not include $\mathbf{0}$, set $\hat{\nu}$ to a sufficiently small positive value.

The minimized CVaR (9), equivalently (27), is decreasing with respect to ν. The minimized CVaR is positive as long as $\nu < \hat{\nu}$, and it is negative as long as $\nu > \hat{\nu}$. When $\mathcal{U}$ of (13) is defined with $\nu > \hat{\nu}$, (9) with $\mathcal{K} = \mathbb{R}^n$ is equivalent to the convex relaxation problem (10).

In machine learning methods, we need to solve (9) many times by changing the size (e.g., ν in the above example) of $\mathcal{U}$ to achieve good prediction performance. Once we get the threshold $\hat{\nu}$, it is enough to solve the convex problem (10) as long as ν is larger than $\hat{\nu}$. In that sense, $\hat{\nu}$ of (28) is more useful than τ of (6) as a criterion of the transferability of (9) to the convex relaxation problem.

6 Numerical results

All the numerical experiments in this paper were performed on an Intel Core i7 2.3 GHz personal computer with 8GB of physical memory using Matlab (R2011b) with IBM ILOG CPLEX 12.

6.1 Portfolio optimization of hedge funds

We consider the hedge funds portfolio optimization problem discussed in Sect. 5.1.1:

$$\begin{aligned} &\min_{\boldsymbol{v}} \ \mu(\tilde{\boldsymbol{z}}^\top \boldsymbol{v}), \\ &\text{s.t.} \ \ \|\boldsymbol{v}\|_1 = c, \ \boldsymbol{e}^\top \boldsymbol{v} = 1, \ \ -\gamma \le \sum_{i \in G_k} v_i \le \gamma, \ \ k = 1, \ldots, K, \end{aligned} \tag{29}$$

for $c > 0$, $\gamma > 0$. Let G_k be the set of indices for each sector, $k = 1, \ldots, K$. Note that the parameter $\gamma \ge 1/K$ since the constraints of (29) lead to

$$1 = \sum_{k=1}^{K} \sum_{i \in G_k} v_i \le \sum_{k=1}^{K} \gamma = K\gamma.$$

We evaluate the effect of constraints $\|\boldsymbol{v}\|_1 = c$ and $-\gamma \le \sum_{i \in G_k} v_i \le \gamma$, $\forall k$, in this section.

We randomly selected n stocks from the stocks listed in the *Nikkei 225 index* and used historical asset return data for $\tilde{\boldsymbol{z}}$ from monthly and weekly return data of those stocks. The monthly data set consists of returns of n companies during the 270 consecutive months between May 1987 and October 2009, whereas the weekly data set consists of returns of the n companies during the 1178 consecutive weeks from April 12, 1987 to November 1, 2009. Each company has an industry code (from 1 to 17) determined by SICC. If n stocks are chosen from all industries, K equals to 17.

We designed a portfolio by a global solution to ℓ_1-norm portfolio model (29) using historical data (time window of m length) for $\tilde{\boldsymbol{z}}$. Here we set $\mathcal{Z} = \{\hat{\boldsymbol{z}}_t, \ldots, \hat{\boldsymbol{z}}_{t+m-1}\}$ and define the matrix form $\boldsymbol{Z} = [\hat{\boldsymbol{z}}_t, \ldots, \hat{\boldsymbol{z}}_{t+m-1}]$ for $m = 120$ (10 years) consecutive periods from the monthly data set or for $m = 150$ (almost 3 years) consecutive periods from the weekly data set by shifting the starting time t of the window. The dimension of $\hat{\boldsymbol{z}}_t$ is (%/month) or (%/week) for all t. Let $\widehat{\boldsymbol{v}}_t$ be a decision vector learned from the training (in-sample) dataset, $\hat{\boldsymbol{z}}_t, \ldots, \hat{\boldsymbol{z}}_{t+m-1}$, of length m. We evaluated the test (out-of-sample) expected return $\hat{\boldsymbol{z}}_{t+m}^\top \widehat{\boldsymbol{v}}_t$ for the next-step sample $\hat{\boldsymbol{z}}_{t+m}$. This procedure was repeated for $t = 1, \ldots, \overline{T}$ ($\overline{T} = 150$ for the monthly data set and $\overline{T} = 1028$ for the weekly data set). The mean of the training expected returns (training.ER) over $\overline{T}$ of time was computed as

$$\frac{1}{\overline{T}} \sum_{t=1}^{\overline{T}} \left\{ \frac{1}{m} \sum_{s=0}^{m-1} \hat{\boldsymbol{z}}_{t+s}^\top \widehat{\boldsymbol{v}}_t \right\}.$$

The mean of the test expected returns (test.ER):

$$\frac{1}{\overline{T}} \sum_{t=1}^{\overline{T}} \hat{\boldsymbol{z}}_{t+m}^\top \widehat{\boldsymbol{v}}_t$$

was employed as performance measures. We also calculated the measure of risk-adjusted performance known as Sharpe ratio:

$$\mathrm{SR} = (\text{test.ER})/\sigma$$

using the mean of test expected returns and their standard deviation σ. The turnover of the portfolio:

$$\text{turnover} = \frac{1}{\overline{T}-1} \sum_{t=2}^{\overline{T}} \sum_{j=1}^{n} \left| \widehat{v}_{t,j} - \frac{1+\hat{z}_{t+m,j}}{1+\hat{\boldsymbol{z}}_{t+m}^\top \widehat{\boldsymbol{v}}_{t-1}} \widehat{v}_{t-1,j} \right|$$

is another performance measure. $\frac{1+\hat{z}_{t+m,j}}{1+\hat{\boldsymbol{z}}_{t+m}^\top \widehat{\boldsymbol{v}}_{t-1}} \widehat{v}_{t-1,j}$ indicates the portfolio weight before rebalancing but at time t, whereas $\widehat{v}_{t,j}$ is the desired portfolio weight at t (after rebalancing). The portfolio turnover measures how frequently assets in a portfolio are bought and sold. The turnover is preferably small, while large test.ER and large SR are preferred.

6.1.1 Minimizing CVaR

Table 1 compares proposed model (29) to "NormConst" and "NoNormConst" using the CVaR risk measure μ for monthly historical dataset consisting of $n = 10$, 20 and 30 assets. "NormConst" means the portfolio model where industry constraints $-\gamma \leq \sum_{i \in G_k} v_i \leq \gamma$, $\forall k$, are ignored from (29), and "NoNormConst" means the

Table 1 CVaR risk measure minimization for Nikkei monthly dataset

	nonconv.ratio	training.ER	test.ER	var.	SR	turnover
$n = 10$						
NoNormConst	–	0.3403	−0.0034	30.3403	−0.0006	0.1903
NormConst	0.2133	0.3242	0.1622	**26.6234**	0.0314	0.1326
Proposed (29)	0.8067	0.4282	**0.4937**	44.9360	**0.0737**	**0.0969**
$n = 20$						
NoNormConst	–	0.5675	0.1176	39.4722	0.0187	0.3564
NormConst	0	0.4187	0.1195	**21.4081**	0.0258	0.1537
Proposed (29)	0	0.4378	**0.4607**	31.7215	**0.0818**	**0.1188**
$n = 30$						
NoNormConst	–	0.8645	−0.3459	59.6936	−0.0448	0.7309
NormConst	0	0.5093	0.1525	**22.7495**	0.0320	0.1844
Proposed (29)	0	0.4646	**0.3356**	25.7000	**0.0662**	**0.1273**

Table 2 CVaR risk measure minimization for Nikkei weekly dataset

	nonconv.ratio	training.ER	test.ER	var.	SR	turnover
$n = 10$						
NoNormConst	–	0.0786	−0.0594	7.0436	−0.0224	0.0894
NormConst	0.1936	0.0684	−0.0522	**6.8235**	−0.0200	0.0770
Proposed (29)	0.2607	0.0792	**0.0160**	9.5446	**0.0052**	**0.0478**
$n = 15$						
NoNormConst	–	0.1029	−0.0703	7.4985	−0.0257	0.1626
NormConst	0.0010	0.0736	−0.0743	**6.6197**	−0.0289	0.0960
Proposed (29)	0.0068	0.0881	**0.0365**	8.8006	**0.0123**	**0.0557**
$n = 20$						
NoNormConst	–	0.1394	−0.0788	7.5242	−0.0287	0.2228
NormConst	0	0.0917	−0.0659	**6.4888**	−0.0259	0.1051
Proposed (29)	0	0.0886	**0.0229**	8.4467	**0.0079**	**0.0690**

model where the norm constraint $\|\boldsymbol{v}\|_1 = c$ is ignored from (29) as well as the above industry constraints. The types K of industries of $n = 10, 20$ and 30 assets were 8, 11 and 14, respectively. We set ν in $\mathcal{U}$ of (13) for the CVaR risk measure to $\nu = 0.2$ and γ of (29) to its almost lower bound as $\gamma = 1/K + \epsilon$ (ϵ is a small value so that $1/K$ was rounded to the second decimal place).

We select c of "NormConst" and (29) during running the algorithm in the following way. Given training (in-sample) dataset (time window of m length), we obtained solutions of each model with different c ($c = 1.1$, 1.2 and 1.3) using first 5/6 of samples and found the suitable parameter value c by the one which attained the maximum expected returns for the remaining 1/6 of samples. The portfolios of "NormConst" and (29) were obtained with such parameter values c. We can easily check the transferabil-

ity of (29) to the convex relaxation problem by comparing c with the norm-threshold τ once we obtain τ by solving the convex optimization (5), i.e., the convex problem constructed by removing the nonconvex constraint $\|\boldsymbol{v}\|_1 = c$ from (29). nonconv.ratio shows the ratio of nonconvex problems that were solved among $\overline{T} = 150$. In the case of $n = 10$, the nonconvex cases of the ℓ_1-norm constraint $\|\boldsymbol{v}\|_1 = c$ occurred for "NormConst" and (29), whereas in $n = 20$ and 30 datasets, the nonconvex cases did not occur. Indeed, the convexity thresholds τ, computed from (6), were sufficiently large for $n = 20$ and 30 datasets, and therefore, the nonconvex constraint was equivalently transformed to $\|\boldsymbol{v}\|_1 \leq c$. When nonconv.ratio equals to 0, "NormConst" is exactly the same as the norm-constraint portfolio optimization model (24).

As the asset size n increases, training.ER becomes larger in all models. This is a reasonable observation because more assets make it possible to attain the best portfolio fit to the given datasets. However, such optimal portfolios with large n assets tend to be overfitting to the given datasets and could not achieve good performance for new return in terms of test.ER. We can also see the similar tendency for three models with $n = 30$. Indeed, our model had the best performance in terms of test.ER, SR and turnover, though other two models achieved larger training.ER especially for $n = 30$.

Table 2 compares proposed model (29) to "NormConst" and "NoNormConst" using the CVaR risk measure μ for weekly historical dataset consisting of $n = 10$, 15 and 20 assets with $K = 8$, 10 and 11 industries, respectively. The parameter settings are exactly the same to those of numerical experiments with monthly datasets except for candidate values of c for "NormConst" and (29): the best c was found among $c = 1.05$, 1.1 and 1.15. As well as numerical experiments for monthly dataset, our model (29) achieved the best performance among three models in terms of test.ER, SR and turnover.

6.1.2 Coherent and non-coherent risk measure minimization

We did similar numerical experiments for our model (29) using a new coherent measures μ based on the Bregman divergence with squared Mahalanobis distance, shown in Example 4. The corresponding uncertainty set $\mathcal{U}$ is given by (16), which is the intersection of the convex hull of data points $\boldsymbol{z}_i$, $i = 1, \ldots, m$, and the ellipsoid with center $\bar{\boldsymbol{z}} = \sum_i \boldsymbol{z}_i / m$ and shape described by the inverse of the covariance matrix $\boldsymbol{\Sigma}_z$ (see (d) in Fig. 1). If we remove the restriction of being the convex hull of data points from $\mathcal{U}$ of (16) and use the resulting set $\widehat{\mathcal{U}}$ of (17) as an uncertainty set $\mathcal{U}$, the risk minimization problem (3) coincides with a classical mean-standard deviation portfolio allocation problem. The mean-standard deviation risk measure is not a coherent risk measure.

The parameter value C in (16) and (17) is fixed to $n/20$ because C is influenced by the dimension n, different from ν of CVaR. We used the same parameter setting for γ and c of (29) as used in the CVaR minimized portfolio model.

Tables 3 and 4 show the performance of portfolio allocation models based on the Bregman divergence (coh.) and minimizing the mean-standard deviation (msd) for Nikkei monthly dataset and Nikkei weekly dataset, respectively. Similarly to the CVaR minimized portfolio model, the model (29) using the Bregman divergence achieved the best performance among three models in terms of test.ER, SR and turnover. Moreover,

Table 3 The coherent risk minimization based on Bregman divergence (coh.) and the mean-standard deviation (msd) minimization for Nikkei monthly dataset

	nonconv.ratio	training.ER	test.ER	var.	SR	turnover
$n = 10$						
coh.						
NoNormConst	–	0.3929	0.1852	33.5758	0.0320	0.1388
NormConst	0.4600	0.3809	0.1551	**28.5189**	0.0290	0.1162
Proposed (29)	0.7733	0.4531	**0.6044**	48.6692	**0.0866**	**0.0993**
msd						
NoNormConst	–	0.3929	0.1852	33.5758	0.0320	0.1388
NormConst	0.5000	0.3675	0.1613	30.6813	0.0291	0.1488
Proposed (29)	0.7733	0.4615	0.5921	49.1772	0.0844	0.1014
$n = 20$						
coh.						
NoNormConst	–	0.5451	0.1516	28.5356	0.0284	0.2395
NormConst	0	0.4179	0.2202	**22.4922**	0.0464	0.1035
Proposed (29)	0	0.4242	**0.4839**	29.4037	**0.0892**	**0.0886**
msd						
NoNormConst	–	0.5459	0.1364	28.1943	0.0257	0.2227
NormConst	0	0.4219	0.1295	23.8368	0.0265	0.1793
Proposed (29)	0	0.4427	0.4210	28.7550	0.0785	0.0984
$n = 30$						
coh.						
NoNormConst	–	0.7241	−0.5472	101.4126	−0.0543	1.0043
NormConst	0	0.4376	0.2149	19.3609	0.0488	0.1421
Proposed (29)	0	0.4424	**0.2820**	24.2553	**0.0573**	0.1149
msd						
NoNormConst	–	0.5204	0.0085	26.1599	0.0017	0.3150
NormConst	0	0.3767	0.1287	**18.7299**	0.0297	0.1973
Proposed (29)	0	0.4585	0.2370	23.0839	0.0493	**0.1071**

we confirm from these tables that the coherent risk measure based on the Bregman divergence is appropriate measure compared to non-coherent risk measure based on the mean-standard deviation for portfolio optimization. Indeed, the coherent risk measure minimization achieved better performance than the non-coherent one except for the turnover and var. in the case of $n = 30$ for Nikkei monthly dataset.

6.2 Binary classification model

In the classification problem setting, the problem (1) minimizing the CVaR risk measure is equivalent to Eν-SVM (27). Eν-SVM (27) is difficult to solve exactly because of the nonconvex constraint $\|\boldsymbol{w}\|_2 = 1$. To find a local optimal solution for Eν-SVM, [19]

Table 4 The coherent risk minimization based on Bregman divergence (coh.) and the mean-standard deviation (msd) minimization for Nikkei weekly dataset

	nonconv.ratio	training.ER	test.ER	var.	SR	turnover
$n = 10$						
coh.						
NoNormConst	–	0.0983	−0.0700	6.7167	−0.0270	0.0690
NormConst	0.2033	0.0879	−0.0511	**6.4965**	−0.0200	0.0595
Proposed (29)	0.4484	0.0842	**0.0077**	9.3668	**0.0025**	**0.0428**
msd						
NoNormConst	–	0.0983	−0.0700	6.7167	−0.0270	0.0690
NormConst	0.2539	0.0987	−0.0636	6.6868	−0.0246	0.0704
Proposed (29)	0.4446	0.0879	0.0014	9.4676	0.0005	0.0461
$n = 15$						
coh.						
NoNormConst	–	0.1124	−0.0734	6.6039	−0.0285	0.1007
NormConst	0.0019	0.0790	−0.0530	**6.1997**	−0.0213	0.0649
Proposed (29)	0.0107	0.0922	**0.0220**	8.8630	**0.0074**	**0.0491**
msd						
NoNormConst	–	0.1124	−0.0734	6.6039	−0.0285	0.1007
NormConst	0.0136	0.1024	−0.0638	6.4990	−0.0250	0.0921
Proposed (29)	0.0097	0.1087	0.0170	8.7689	0.0057	0.0600
$n = 20$						
coh.						
NoNormConst	–	0.1319	−0.0537	6.8060	−0.0206	0.1316
NormConst	0	0.0904	−0.0375	**6.1941**	−0.0150	0.0703
Proposed (29)	0	0.0883	**0.0233**	8.3588	**0.0081**	**0.0624**
msd						
NoNormConst	–	0.1319	−0.0537	6.8060	−0.0206	0.1316
NormConst	0	0.1182	−0.0362	6.6094	−0.0141	0.1122
Proposed (29)	0	0.1076	0.0178	8.0960	0.0062	0.0753

proposed an iterative local search algorithm (modified algorithm of [12]) that solves linearized problems where the quadratic equality constraint $\boldsymbol{w}^\top \boldsymbol{w} = 1$ is replaced by $\bar{\boldsymbol{w}}_k^\top \boldsymbol{w} = 1$ constructed at feasible solutions $\bar{\boldsymbol{w}}_k$, and finds new feasible solutions $\bar{\boldsymbol{w}}_{k+1}$ by projecting the resulting optimal solutions onto the quadratic surface, $\boldsymbol{w}^\top \boldsymbol{w} = 1$. [20] moreover incorporated the iterative local search algorithm in cutting plane scheme in order to find a global optimal solution of Eν-SVM (27). The global search algorithm repeatedly finds a local solution in the remaining feasible region and removes the region consisting of the local solution and its neighbourhood. However, the global search algorithm needs large computational time especially when the problem size (m, n) becomes large.

In this section, we solve ℓ_1 and ℓ_∞-norm problems ((21) and (22), respectively) corresponding to $\mathcal{U}$ of (13) for the CVaR risk measure. Using UCI repository datasets [4], we compare prediction performance of global optimal solutions of ℓ_1-norm and ℓ_∞-norm problems to local optimal solutions of Eν-SVM (27) found by the local search algorithm [19]. As a measure of prediction performance, we use average test (out-of-sample) errors [%] for 3-fold cross validation.

Figure 2 shows average test errors and average computation times with respect to the parameter ν for four datasets: `liver-disorders` ($(n-1)=6, m=345$), `diabetes` ($(n-1)=8, m=768$), `heart` ($(n-1)=13, m=270$) and `breast-cancer` ($(n-1)=10, m=683$). Remind that inputs for (27) and its ℓ_1/ℓ_∞ variants are sample vectors $\boldsymbol{x}_i \in \mathbb{R}^{n-1}$ and $y_i \in \{+1, -1\}$, $i = 1, \ldots, m$.

As discussed in Remark 2, we can easily check the transferability of Eν-SVM (27) to the convex relaxation, ν-SVM, once we get the threshold $\hat{\nu}$ by solving (28); it is sufficient to solve ν-SVM as long as $\nu > \hat{\nu}$. The green vertical line means that at least one of the three problems (3) constructed in the cross-validation was essentially nonconvex, i.e., the green line indicates the threshold for convexity, largest $\hat{\nu}$ among 3-fold cross validation.

The figures in the left column of Fig. 2 show that either of global solutions of ℓ_1-norm and ℓ_∞-norm models achieved better prediction performance than a local solution of ℓ_2-norm Eν-SVM (27) in most values of ν. We can see that ℓ_2-norm model achieved intermediate prediction performance under three different types of norm constraints. This implies that computing global solutions for ℓ_1-norm and ℓ_∞-norm models may be enough to find good performance classifiers.

Classifiers with good prediction performance were found by convex problems with $\nu > \hat{\nu}$ for three UCI datasets. However, for liver-disorders dataset, nonconvex cases of all three models performed well. The computational time in the right column shows that in convex cases ($\nu > \hat{\nu}$), linear optimization formulations of ℓ_1-norm or ℓ_∞-norm problems were solved faster than the quadratic optimization formulation of ℓ_2-norm problem which has the quadratic constraint $\boldsymbol{w}^\top \boldsymbol{w} \leq 1$. On the other hand, in nonconvex cases ($\nu \leq \hat{\nu}$), MIO formulations of ℓ_1-norm or ℓ_∞-norm problems required more computational times than the iterative local search algorithm [19] that finds a local optimum solution. However, within reasonable computational time, we could obtain global solutions of ℓ_1-norm or ℓ_∞-norm problems by those MIO formulations.

7 Conclusion

We have proposed minimizing a coherent risk measure under a norm equality constraint with the use of robust optimization formulation. Not only well-known coherent risk measures but also a new coherent risk measure was investigated by setting a new uncertainty set with the use of a Bregman divergence. The norm equality constraint itself has a practical meaning or plays a role to prevent a meaningless solution, the zero vector, in the context of portfolio optimization or binary classification in machine learning, respectively. We have proposed to use an ℓ_1 or ℓ_∞-norm constraint for achieving a global optimal solution for the nonconvex optimization problems with the use of a mixed integer optimization formulation. Numerical experiments confirmed

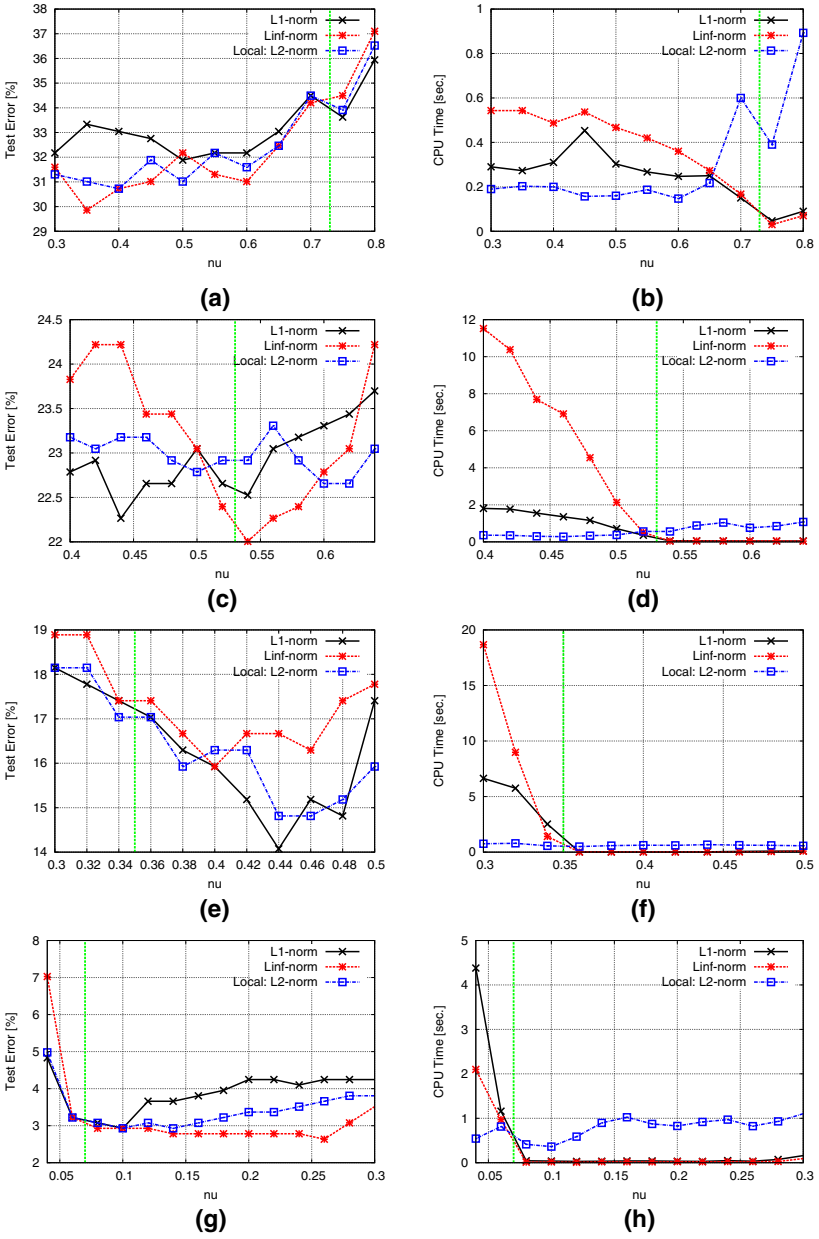


Fig. 2 Classification results for UCI datasets. The *left column* and the *right column* respectively indicate average test (out-of-sample) errors [%] and average CPU time [s] for threefold cross validation with respect to ν. *Each row* shows numerical results for `liver-disorders` ($(n-1)=6, m=345$), `diabetes` ($(n-1)=8, m=768$), `heart` ($(n-1)=13, m=270$) and `breast-cancer` ($(n-1)=10, m=683$) from UCI repository dataset [4]. The left-hand side of the *green vertical line* in each figure indicates that the problem (3) is essentially nonconvex, i.e., $\nu \leq \hat{\nu}$. **a** Test error (liver-disorders), **b** CPU time (liver-disorders), **c** test error (diabetes), **d** CPU time (diabetes), **e** test error (heart), **f** CPU time (heart), **g** test error (breast-cancer), **h** CPU time (breast-cancer)

that the ℓ_1 or ℓ_∞-norm equality constraint as well as coherent risk measures works effectively in portfolio optimization and binary classification.

There are a lot of possibilities for uncertainty set $\mathcal{U}$ to improve our binary classification model and portfolio optimization models shown in the numerical experiments. For portfolio optimization, we tested two models using coherent risk measures; CVaR and Bregman divergence. The portfolio model based on Bregman divergence seems to perform better than the one based on CVaR for Nikkei weekly dataset with $n = 20$, but it may not be true for the other financial datasets. As a future work, we want to consider a method for constructing a suitable uncertainty set $\mathcal{U}$ that fits to a dataset in each application.

References

1. Artzner, P., Delbaen, F., Eber, J., Heath, D.: Coherent measures of risk. Math. Financ. **9**, 203–228 (1999)
2. Bertsimas, D., Brown, D.: Constructing uncertainty sets for robust linear optimization. Oper. Res. **57**(6), 1483–1495 (2009)
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Berlin (2006)
4. Blake, C.L., Merz, C.J.: UCI repository of machine learning databases (1998)
5. Brodie, J., Daubechies, I., De Molc, C., Giannone, D., Loris, I.: Sparse and stable markowitz portfolios. PNAS **106**, 12267–12272 (2009)
6. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
7. DeMiguel, V., Garlappi, L., Nogales, F.J., Uppal, R.: A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. Manag. Sci. **55**, 798–812 (2009)
8. Fischer, T.: Examples of coherent risk measures depending on one-sided moments. In: Working Paper (2001)
9. Gotoh, J., Takeda, A.: On the role of norm constraints in portfolio selection. Comput. Manag. Sci. **8**(4), 323–353 (2011)
10. Gotoh, J., Takeda, A., Yamamoto, R.: Interaction between financial risk measures and machine learning methods. Comput. Manag. Sci. **11**, 365–402 (2014)
11. Natarajan, K., Pachamanova, D., Sim, M.: Constructing risk measures from uncertainty sets. Oper. Res. **57**(5), 1129–1141 (2009)
12. Perez-Cruz, F., Weston, J., Hermann, D.J.L., Schölkopf, B.: Extension of the ν-SVM range for classification. In: Advances in Learning Theory: Methods, Models and Applications, pp. 179–196. IOS Press, Amsterdam (2003)
13. Rockafellar, R.T., Uryasev, S.: Optimization of conditional value-at-risk. J. Risk **2**, 21–41 (2000)
14. Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. J. Bank. Financ. **26**(7), 1443–1472 (2002)
15. Ruszczynski, A., Shapiro, A.: Optimization of convex risk functions. Math. Oper. Res. **31**(3), 433–452 (2006)
16. Schölkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge, MA (2002)
17. Schölkopf, B., Smola, A., Williamson, R., Bartlett, P.: New support vector algorithms. Neural Comput. **12**(5), 1207–1245 (2000)
18. Stubbs, R.A., Vandenbussche, D.: Constraint attribution. J. Portf. Manag. **36**(4), 48–59 (2010)
19. Takeda, A., Sugiyama, M.: Nu-support vector machine as conditional value-at-risk minimization. In: Proceedings of International Conference on Machine Learning, pp. 1056–1063 (2008)
20. Takeda, A., Sugiyama, M.: On generalization performance and non-convex optimization of extended nu-support vector machine. New Gener. Comput. **27**, 259–279 (2009)
21. Takeda, A., Mitsugi, H., Kanamori, T.: A unified classification model based on robust optimization. Neural Comput. **25**(3), 759–804 (2013)
22. Xu, H., Caramanis, C., Mannor, S., Yun S.: Risk sensitive robust support vector machines. In: IEEE Conference on CDC, pp. 4655–4661, (2009)