European Journal of Operational Research 316 (2024) 46–60

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Continuous optimization

A distributionally robust chance-constrained kernel-free quadratic surface support vector machine

Fengming Lin [a], Shu-Cherng Fang [a], Xiaolei Fang [a], Zheming Gao [b,d,*], Jian Luo [c]

[a] Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina, 27695, USA
[b] College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning, 110819, China
[c] International Business School, Hainan University, Haikou, Hainan, 570228, China
[d] Yunnan Key Laboratory of Service Computing, Kunming, Yunnan, 650221, China

ARTICLE INFO

Keywords:
Data science
Kernel-free support vector machine
Robust classification
Distributionally robust optimization
Chance-constrained optimization

ABSTRACT

This paper studies the problem of constructing a robust nonlinear classifier when the data set involves uncertainty and only the first- and second-order moments are known a priori. A distributionally robust chance-constrained kernel-free quadratic surface support vector machine (SVM) model is proposed using the moment information of the uncertain data. The proposed model is reformulated as a semidefinite programming problem and a second-order cone programming problem for efficient computations. A geometric interpretation of the proposed model is also provided. For commonly used data without prescribed uncertainty, a cluster-based data-driven approach is introduced to retrieve the hidden moment information that enables the proposed model for robust classification. Extensive computational experiments using synthetic and public benchmark data sets with or without uncertainty involved support the superior performance of the proposed model over other state-of-the-art SVM models, particularly when the data sets are massive and/or imbalanced.

1. Introduction

Support Vector Machines (SVMs) are often used for classification in supervised machine learning. Given a set of N data points $\{(\boldsymbol{x}^i, y^i) | \boldsymbol{x}^i \in \mathbb{R}^n,\ y^i \in \{-1, 1\}, i = 1, \ldots, N\}$, a linear soft support vector machine (LSSVM) can be represented as the following linearly constrained convex quadratic programming problem (Cortes & Vapnik, 1995):

$$\begin{aligned} \min\quad & \tfrac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i \\ s.t.\quad & y^i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i + b\right) \geqslant 1 - \xi_i,\ i = 1, \ldots, N, \\ & \boldsymbol{w} \in \mathbb{R}^n,\ b \in \mathbb{R},\ \boldsymbol{\xi} \in \mathbb{R}_+^N, \end{aligned} \tag{LSSVM}$$

where $C > 0$ is a given parameter. The variables of $\boldsymbol{w}$ and b determine a separation hyperplane $H(\boldsymbol{w}, b) \triangleq \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0\}$, and the slack vector $\boldsymbol{\xi}$ introduces a "soft margin" to accommodate the data that are not linearly separable. For nonlinear classification, the data can be lifted to a higher dimensional space for linear separation using a feature map $\phi : \mathbb{R}^n \to \mathbb{R}^l$ with $l > n$. In this case, we consider the following optimization problem:

$$\begin{aligned} \min\quad & \tfrac{1}{2}\|\boldsymbol{v}\|_2^2 + C\sum_{i=1}^{N}\xi_i \\ s.t.\quad & y^i\left(\boldsymbol{v}^{\mathrm{T}}\phi(\boldsymbol{x}^i) + d\right) \geqslant 1 - \xi_i,\ i = 1, \ldots, N, \\ & \boldsymbol{v} \in \mathbb{R}^l,\ d \in \mathbb{R},\ \boldsymbol{\xi} \in \mathbb{R}_+^N. \end{aligned} \tag{KSSVM}$$

The kernel trick is used to solve the dual problem of (KSSVM) by introducing the kernel function $K(\boldsymbol{x}^i, \boldsymbol{x}^j) \triangleq \phi(\boldsymbol{x}^i)^{\mathrm{T}}\phi(\boldsymbol{x}^j)$ (Zhou, 2021). Commonly used kernel functions $K(\cdot, \cdot)$ include the polynomial kernel and radial basis function kernel (i.e., Gaussian kernel). Considering the drawbacks of selecting a proper kernel function (Jiménez-Cordero, Morales, & Pineda, 2021) and adjusting its embedded parameters, some kernel-free nonlinear SVMs have recently been proposed (Dagher, 2008; Luo, Fang, Deng, & Guo, 2016; Luo, Yan, & Tian, 2020). One representative model is the following quadratic surface support vector machine (QSSVM):

$$\begin{aligned} \min\quad & \sum_{i=1}^{N}\|\boldsymbol{M}\boldsymbol{x}^i + \boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i \\ s.t.\quad & y^i\left(\tfrac{1}{2}(\boldsymbol{x}^i)^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}^i + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i + b\right) \geqslant 1 - \xi_i,\ i = 1, \ldots, N, \\ & \boldsymbol{M} \in \mathbb{S}^n,\ \boldsymbol{w} \in \mathbb{R}^n,\ b \in \mathbb{R},\ \boldsymbol{\xi} \in \mathbb{R}_+^N, \end{aligned} \tag{QSSVM}$$

where $\mathbb{S}^n$ is the set of n-dimensional symmetric matrices; $\boldsymbol{M}$, $\boldsymbol{w}$, and b determine a separation quadratic surface $Q(\boldsymbol{M}, \boldsymbol{w}, b) \triangleq \{\boldsymbol{x} \in \mathbb{R}^n | \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} + b = 0\}$. The (QSSVM) model has been extended to a kernel-free quartic surface SVM model by utilizing the double well potential function of degree four (Gao, Fang, Luo, & Medhin, 2021).

While both the kernel-based and kernel-free nonlinear SVMs have achieved promising performance in some real-world applications, their

* Corresponding author at: College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning, 110819, China.
E-mail address: gaozheming@ise.neu.edu.cn (Z. Gao).

https://doi.org/10.1016/j.ejor.2024.02.022
Received 16 February 2023; Accepted 18 February 2024
Available online 22 February 2024

classification performances still need to be investigated when uncertainty is involved in the training data. For instance, the classification task on benign and malignant tumors (Bertsimas, Dunn, Pawlowski, & Zhuo, 2019), whose training data includes features derived from digitized images, such as cell nuclei radius, texture, and symmetry. Even though these features are precisely measured, the existence of image noise and measurement inaccuracy may yield data uncertainty and affect the classification accuracy. In addition to medical applications, challenges raised by data uncertainties are urgent to be resolved in other fields, including battery failure detection (Luo, Fang, Deng, & Tian, 2022) and biological gene expression (Ben-Tal, Bhadra, Bhattacharyya, & Nath, 2011). In datasets requiring imputation for missing data (Shivaswamy, Bhattacharyya, & Smola, 2006), additional uncertainties are introduced. Given the presence of data uncertainty within real-world applications, neglecting to recognize the uncertainty might lead to a substantial decline in classification performance (Goldfarb & Iyengar, 2003).

Recent studies indicate that classifiers explicitly addressing uncertainty in the training data outperform those ignoring such information (Wang, Fan, & Pardalos, 2018). This paper introduces a novel maximum-margin nonlinear SVM resilient to data uncertainty. It handles the underlying data uncertainty by utilizing moment information instead of the distributional assumptions on data.

1.1. Relevant works

Optimization under uncertainty has been addressed by several complementary modeling paradigms that differ mainly in the representation of uncertainty. SVM models applying robust optimization techniques are developed for applications whose data points are fluctuating within an uncertain set, specified by the l_p-norm uncertainty (Trafalis & Gilbert, 2006), ellipsoidal uncertainty (Bhattacharyya, Grate, Jordan, Ghaoui, & Mian, 2004), and others (Singla, Ghosh, & Shukla, 2020; Wang & Pardalos, 2014). Applying robust optimization in a principled way of uncertain data, Bertsimas et al. (2019) investigate the SVMs, logistic regression, and decision trees, among which the robust SVM performed the best. Nonetheless, the robust models generally tend to be on the conservative side since they ignore the hidden distribution information embedded in the data sets.

Consider a set of data points with uncertain inputs following some underlying probability distributions F_i, i.e., $\tilde{\boldsymbol{x}}^i \sim F_i$, for $i = 1, \ldots, N$. For a given tolerance level $0 < \epsilon < 1$, a chance constraint at the point $(\tilde{\boldsymbol{x}}^i, y^i)$,

$$\mathbb{P}_{F_i}\left\{y^i\left(\boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b\right) \leqslant 1 - \xi_i\right\} \leqslant \epsilon,$$

can be used to ensure that the probability of misclassifying $\tilde{\boldsymbol{x}}^i$ is no larger than ϵ. The chance-constrained optimization problems are nonconvex and hard to solve in general. In the literature, Peng, Gianpiero, and Zhihua (2023) adopt the sample average approximation method to formulate a mixed integer programming problem for a chance-constrained conic-segmentation SVM with an empirical distribution. In fact, a true distribution is hard to estimate, and even a good estimation may still cause the "optimizer's curse" (Kuhn, Esfahani, Nguyen, & Shafieezadeh-Abadeh, 2019) with discontent performance.

Instead of relying on a single estimate of F_i, a distribution family $\mathcal{D}_i$ could hedge against the uncertainty in data distribution. $\mathcal{D}_i$, which is also known as the ambiguity set (Lin, Fang, & Gao, 2022), consists of probability distributions possessing certain properties of the true distribution F_i. The following distributionally robust chance constraints are developed to ensure that a linear SVM works best in the worst case over $\mathcal{D}_i$:

$$\sup_{F_i \in \mathcal{D}_i} \mathbb{P}_{F_i}\left\{y^i\left(\boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b\right) \leqslant 1 - \xi_i\right\} \leqslant \epsilon,\ i = 1, \ldots, N. \tag{1}$$

When the ambiguity set $\mathcal{D}_i$ is constructed based on the first- and second-order moments, the distributionally robust chance-constrained linear soft SVM (DRC-LSSVM) model is developed with the chance constraints defined by (1). Shivaswamy et al. (2006) adopt the multivariate Chebyshev inequality to derive a second-order cone programming (SOCP) reformulation for the DRC-LSSVM model. Ben-Tal et al. (2011) employ the Bernstein bounds to include richer partial information for constructing a less conservative SOCP reformulation. Wang et al. (2018) derive both semidefinite programming (SDP) and SOCP reformulations for DRC-LSSVM, and they further design a stochastic gradient-based method for improving the computational efficiency in large-scale classification cases (Wang, Fan, & Pardalos, 2017). Considering dependency among the random input points, Khanjani-Shiraz, Babapour-Azar, Hosseini-Nodeh, and Pardalos (2023) propose a robust joint chance-constrained linear SVM. The DRC-LSSVM model has also been applied to different contexts with promising performance, such as data with missing values (Shivaswamy et al., 2006) and semi-supervised classifications (Huang, Song, Gupta, & Wu, 2013). A kernel-free DRC support vector regression (SVR) model is proposed to solve regression problems, which also shows superior performance over other well-established SVR models (Luo et al., 2022). Similar links to supervised training with uncertain data employing distributionally robust optimization under the Wasserstein metric have been investigated for linear SVMs (Ma & Wang, 2021), regression models (Chen & Paschalidis, 2020; Kuhn et al., 2019) and reinforcement learning models (Chen & Paschalidis, 2020).

In summary, the literature indicates that the application of distributionally robust optimization enhances the capability of conventional SVMs in addressing uncertain classification tasks. Previous studies on distributionally robust chance-constrained SVMs demonstrate higher accuracy compared to nominal methods in some cases, but most of them focus on linear SVMs, limiting the applicability to nonlinear classification. Our contribution builds on these efforts by proposing a distributionally robust chance-constrained kernel-free nonlinear SVM, utilizing a quadratic surface for increased flexibility in handling nonlinear data. We compare this approach to state-of-the-art SVMs, assessing the impact of adding robustness to different models and evaluating their performance through computational experiments with balanced, imbalanced, and massive datasets.

1.2. Contributions

To deal with the nonlinear binary classification cases with uncertain data, in this paper, we propose a kernel-free quadratic surface SVM model that considers the distributionally robust chance constraints (2).

$$\sup_{F_i \in \mathcal{D}_i} \mathbb{P}_{F_i}\left\{y^i\left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b\right) \leqslant 1 - \xi_i\right\} \leqslant \epsilon, i = 1, \ldots, N. \tag{2}$$

Certain analytic properties of the proposed model are rigorously investigated. In addition, extensive computational experiments are conducted to validate the effectiveness and efficiency of the proposed model in solving binary classification problems with and without data uncertainty. The main contributions of this paper are summarized as follows.

- We propose a distributionally robust chance-constrained quadratic SVM (DRC-QSSVM) model utilizing the first- and second-order moments embedded in the data set, which characterizes the uncertainty of the classification problem. To the best of our knowledge, it is the first study of utilizing kernel-free nonlinear SVM models to deal with classification problems under data uncertainty. As the quadratic structure of the distributionally robust chance constraints in the proposed model complicates the analysis, we explicitly derive the SDP and the SOCP reformulations of the proposed model for computational efficiency. In addition, a geometric interpretation of the distributionally robust quadratic chance constraints is provided for a better understanding of the proposed model.

- We extend the proposed model to handle commonly used data without uncertainty. To retrieve the moment information needed by the proposed model, a cluster-based data-driven approach is designed. Surprisingly, the proposed model provides higher classification accuracy than the other tested state-of-the-art SVM models in the computational experiments. It strengthens the applicability of the proposed model to real-life applications.
- The computational results verify the classification effectiveness of the proposed DRC-QSSVM model. As a maximum-margin SVM model that can explicitly use moment information to handle input data with uncertainty, the proposed model outperforms the most related DRC-LSSVM model on some synthetic and public benchmark data sets. Also, the results from some extensive computational experiments on massive and imbalanced data sets verify the dominant classification accuracy of the proposed model over other state-of-the-art SVM models. It reveals the significance of the proposed model that reframing a specific problem as one characterized by uncertainty and subsequently addressing the resultant uncertain formulation have the potential to yield remarkably improved outcomes.

The rest of the paper is organized as follows. In Section 2, we propose a distributionally robust chance-constrained quadratic surface support vector machine model for nonlinear classification with uncertain data knowing the first- and second-order moments. The SDP and SOCP reformulations are derived for computational efficiency. A geometric interpretation is also provided to show how the proposed model works on uncertain data. Section 3 presents a data-driven approach for applying the proposed model to classify commonly used data sets without moment information. Synthetic data sets and public benchmark data sets are included in Section 4 for validating the effectiveness and efficiency of the proposed model and comparing the performance of the proposed model with other well-known SVM models. Section 5 concludes the paper.

Notations: In this paper, we use lower-case boldface letters to denote vectors and upper-case boldface letters to denote matrices. Random variables are represented by symbols with tildes, while their realizations are denoted by the same symbols without tildes. $\mathbb{S}^n$ denotes the set of symmetric matrices of dimension n. For any two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^n$, $\boldsymbol{A} \bullet \boldsymbol{B} = Trace(\boldsymbol{A}\boldsymbol{B})$ denotes the trace of the product of $\boldsymbol{A}$ and $\boldsymbol{B}$.

2. Distributionally robust chance-constrained quadratic SVM

This section considers a binary classification problem for uncertain data sets with known first- and second-order moments information and proposes the DRC-QSSVM model. Section 2.1 constructs an ambiguity set based on the first two moments to formalize the proposed model. An equivalent SDP model is derived in Section 2.2. Section 2.3 presents an explicit geometric interpretation of the conceptual chance constraints. An SOCP model for efficient computation is derived in Section 2.4.

2.1. DRC-QSSVM model

In this paper, each uncertain input $\tilde{\boldsymbol{x}}^i$ in a data set of $\{(\tilde{\boldsymbol{x}}^i, y^i) | \tilde{\boldsymbol{x}}^i \in \mathbb{R}^n,\ y^i \in \{-1, 1\}, i = 1, \ldots, N\}$ is considered as a random vector, i.e., $\tilde{\boldsymbol{x}}^i : \Xi_i \to \mathbb{R}$ and $\tilde{\boldsymbol{x}}^i \sim F_i$, for an outcome space Ξ_i and its σ-algebra $\mathcal{F}_i \subseteq 2^{\Xi_i}$, and $F_i : \mathcal{F}_i \to \mathbb{R}$ is a probability measure on $(\Xi_i, \mathcal{F}_i)$, for $i = 1, \ldots, N$. Let $\mathcal{M}(\Xi_i, \mathcal{F}_i)$ denote the space of all probability measures defined on $(\Xi_i, \mathcal{F}_i)$. In this way, $F_i \in \mathcal{M}(\Xi_i, \mathcal{F}_i)$ and they are assumed to be mutually independent for $i = 1, \ldots, N$. In Section 2, the mean $\boldsymbol{\mu}_i \triangleq \mathbb{E}_{F_i}[\tilde{\boldsymbol{x}}_i] \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{\Sigma}_i \triangleq \mathbb{E}_{F_i}[(\tilde{\boldsymbol{x}}_i - \mathbb{E}_{F_i}[\tilde{\boldsymbol{x}}_i])(\tilde{\boldsymbol{x}}_i - \mathbb{E}_{F_i}[\tilde{\boldsymbol{x}}_i])^{\mathrm{T}}] \in \mathbb{S}^n_+$ are particularly assumed to be known. Without loss of generality, $\boldsymbol{\Sigma}_i$ is considered to be positive definite. A moment-based ambiguity set $\mathcal{D}$ is then defined by $\mathcal{D} \triangleq \bigcup_{i=1}^N \mathcal{D}_i(\tilde{\boldsymbol{x}}^i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with

$$\mathcal{D}_i(\tilde{\boldsymbol{x}}^i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \triangleq \left\{ F_i \in \mathcal{M}(\Xi_i, \mathcal{F}_i) \;\middle|\; \begin{array}{l} \mathbb{P}(\tilde{\boldsymbol{x}}^i \in \Xi_i) = 1, \\ \mathbb{E}_{F_i}[\tilde{\boldsymbol{x}}^i] = \boldsymbol{\mu}_i, \\ \mathbb{E}_{F_i}[(\tilde{\boldsymbol{x}}^i - \boldsymbol{\mu}_i)(\tilde{\boldsymbol{x}}^i - \boldsymbol{\mu}_i)^{\mathrm{T}}] = \boldsymbol{\Sigma}_i \end{array} \right\}, \tag{3}$$

and $\mathcal{D}_i$ abbreviates $\mathcal{D}_i(\tilde{\boldsymbol{x}}^i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for $i = 1, \ldots, N$. In our model, we set $\mathbb{R}^n$ as the support set Ξ_i for $i = 1, \ldots, N$.

We aim to determine a quadratic surface $Q(\boldsymbol{M}, \boldsymbol{w}, b) = \{\boldsymbol{x} \in \mathbb{R}^n | \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} + b = 0\}$, where $\boldsymbol{M} \in \mathbb{S}^n$, $\boldsymbol{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, which separates the two classes of points with the maximum margin and bounded misclassification probability with respect to all distributions in $\mathcal{D}$. Adopting the concept of "total approximated relative geometric margins" used in (QSSVM) (Luo et al., 2016), the min–max approach used in robust optimization helps form the following objective function:

$$\min_{\boldsymbol{M}, \boldsymbol{w}, b, \xi} \sup_{F_i \in \mathcal{D}_i} \left\{ \sum_{i=1}^N \mathbb{E}_{F_i} \|\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}\|_2^2 \right\} + C \sum_{i=1}^N \xi_i. \tag{4}$$

Note that for any $F_i \in \mathcal{D}_i$,

$$\begin{aligned} \mathbb{E}_{F_i} \|\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}\|_2^2 &= \mathbb{E}_{F_i}\left[(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i\right] + 2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{M}\mathbb{E}_{F_i}\left[\tilde{\boldsymbol{x}}^i\right] + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} \\ &= \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i + (\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}) \bullet \boldsymbol{\Sigma}_i + 2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} \\ &= \|\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}\|_2^2 + \|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\|_F^2, \end{aligned} \tag{5}$$

where $\|\cdot\|_F$ is the Frobenius norm. When $\{F_i\}_i$ are mutually independent, the supremum and summation operations are exchangeable, and consequently, $\sup_{F_i \in \mathcal{D}_i}\{\sum_{i=1}^N \mathbb{E}_{F_i}\|\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}\|_2^2\} = \sum_{i=1}^N \sup_{F_i \in \mathcal{D}_i} \mathbb{E}_{F_i}\|\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}\|_2^2 = \sum_{i=1}^N \{\|\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}\|_2^2 + \|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\|_F^2\}$. The first term of the results in (5) is the approximated relative geometrical margin at the mean vector $\boldsymbol{\mu}_i$ (Luo et al., 2016), and the second term is similar to the "G-margin" defined in Gao et al. (2021). In general, the total relative geometrical margin dominates the G-margin. And the second term may serve as a regularization term that shapes the target quadratic surface. Extensive computational experiments indicate such a regularization term can be neglected without changing much of the final classifier (similar results showed by Luo et al. (2016)). Moreover, to avoid the computational difficulty induced by $\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\|_F^2$, we omit this term in the objective function. A robust classifier $Q(\boldsymbol{M}, \boldsymbol{w}, b)$ could bound the misclassification probability by ϵ $(0 < \epsilon < 1)$ employing the distributionally robust chance constraints (2). Consequently, we propose the following distributionally robust chance-constrained quadratic SVM model:

$$\begin{aligned} \min \quad & \sum_{i=1}^N \|\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}\|_2^2 + C\sum_{i=1}^N \xi_i \\ s.t. \quad & \sup_{F_i \in \mathcal{D}_i} \mathbb{P}_{F_i}\left\{ y^i\left(\tfrac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b\right) \leqslant 1 - \xi_i \right\} \leqslant \epsilon,\ i = 1, \ldots, N, \\ & \boldsymbol{M} \in \mathbb{S}^n,\ \boldsymbol{w} \in \mathbb{R}^n,\ b \in \mathbb{R},\ \boldsymbol{\xi} \in \mathbb{R}^N_+. \end{aligned} \tag{DRC-QSSVM}$$

It is often the case that the ambiguous chance constraints are hard to solve directly, not to mention the nonlinear functions used in (DRC-QSSVM). For $i = 1, \ldots, N$, let

$$\begin{aligned} \mathcal{V}_i \triangleq \Bigg\{ (\boldsymbol{M}, \boldsymbol{w}, b) \in \mathbb{S}^n \times \mathbb{R}^n \times \mathbb{R} \;\Bigg|\; & \sup_{F_i \in \mathcal{D}_i} \mathbb{P}_{F_i}\left\{ y^i\left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b\right) \right. \\ & \left. \leqslant 1 - \xi_i \right\} \leqslant \epsilon \Bigg\} \end{aligned} \tag{6}$$

denote the feasible sets of (DRC-QSSVM). We shall demonstrate that for any i, $\mathcal{V}_i$ has tractable SDP and SOCP representations for efficient computations in later subsections.

2.2. SDP reformulation of (DRC-QSSVM)

We first show an equivalent SDP reformulation of (DRC-QSSVM) in the next theorem.

Theorem 2.1. *For the ambiguity set $\mathcal{D}$ defined by (3), (DRC-QSSVM) can be equivalently reformulated as the following SDP problem:*

$$\begin{array}{lll}
\min & \sum_{i=1}^{N} \|\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i & \\
s.t. & \beta_i - \frac{1}{\epsilon}\boldsymbol{\Gamma}_i \bullet \boldsymbol{R}_i \geqslant 0, & i=1,\ldots,N,\\
& \boldsymbol{R}_i + \begin{bmatrix} \frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w} \\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^i b + \xi_i - 1 - \beta_i \end{bmatrix} \succeq 0, & i=1,\ldots,N,\\
& \boldsymbol{R}_i \succeq 0, & i=1,\ldots,N,\\
& \boldsymbol{M}\in\mathbb{S}^n,\ \boldsymbol{w}\in\mathbb{R}^n,\ b\in\mathbb{R},\ \boldsymbol{\xi}\in\mathbb{R}^N_+,\ \boldsymbol{\beta}\in\mathbb{R}^N,\ \boldsymbol{R}_i\in\mathbb{S}^{n+1}, & i=1,\ldots,N,
\end{array} \tag{7}$$

where $\boldsymbol{\Gamma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^{\mathrm{T}} & 1 \end{bmatrix}$ denotes the second-order moment matrix.

Proof. For $i=1,\ldots,N$, we denote the indicator functions by

$$\mathbb{1}_{A_i}(\tilde{\boldsymbol{x}}^i) = \begin{cases} 1, & \text{if } \tilde{\boldsymbol{x}}^i \in A_i, \\ 0, & \text{if } \tilde{\boldsymbol{x}}^i \notin A_i, \end{cases}$$

where $A_i \triangleq \{\tilde{\boldsymbol{x}}^i \in \Xi_i | y^i\left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b\right) \leqslant 1-\xi_i\}$. Then $\varphi_i \triangleq \sup_{F_i\in\mathcal{D}_i}\mathbb{P}_{F_i}\{y^i(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b) \leqslant 1-\xi_i\} = \sup_{F_i\in\mathcal{D}_i}\mathbb{E}_{F_i}[\mathbb{1}_{A_i}(\tilde{\boldsymbol{x}}^i)]$, which can be obtained by solving the following problem:

$$\sup \int_{\Xi_i} \mathbb{1}_{A_i}(\tilde{\boldsymbol{x}}^i) dF_i(\tilde{\boldsymbol{x}}^i) \tag{8a}$$

$$s.t. \int_{\Xi_i} dF_i(\tilde{\boldsymbol{x}}^i) = 1, \tag{8b}$$

$$\int_{\Xi_i} \tilde{\boldsymbol{x}}^i dF_i(\tilde{\boldsymbol{x}}^i) = \boldsymbol{\mu}_i, \tag{8c}$$

$$\int_{\Xi_i} (\tilde{\boldsymbol{x}}^i - \boldsymbol{\mu}_i)(\tilde{\boldsymbol{x}}^i - \boldsymbol{\mu}_i)^{\mathrm{T}} dF_i(\tilde{\boldsymbol{x}}^i) = \boldsymbol{\Sigma}_i, \tag{8d}$$

$$F_i \in \mathcal{M}(\Xi_i, \mathcal{F}_i).$$

Notice that constraint (8d) is equivalent to $\int_{\Xi_i}\tilde{\boldsymbol{x}}^i(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}dF_i(\tilde{\boldsymbol{x}}^i) = \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}$, and the difficulty of this problem can be circumvented by using the duality theory involving moment information. Let $r_i\in\mathbb{R}$, $\boldsymbol{p}_i\in\mathbb{R}^n$ and $\boldsymbol{Q}_i\in\mathbb{S}^n$ be the dual variables corresponding to (8b), (8c) and (8d), respectively. Then the dual problem of (8) becomes

$$\begin{array}{ll}
\inf & (\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}) \bullet \boldsymbol{Q}_i + \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{p}_i + r_i \\
s.t. & (\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{Q}_i\tilde{\boldsymbol{x}}^i + (\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{p}_i + r_i \geqslant \mathbb{1}(\tilde{\boldsymbol{x}}^i),\ \forall\ \tilde{\boldsymbol{x}}^i\in\Xi_i, \\
& \boldsymbol{Q}_i\in\mathbb{S}^n,\ \boldsymbol{p}_i\in\mathbb{R}^n,\ r_i\in\mathbb{R}.
\end{array} \tag{9}$$

Let $\boldsymbol{N}_i = \begin{bmatrix} \boldsymbol{Q}_i & \frac{1}{2}\boldsymbol{p}_i \\ \frac{1}{2}\boldsymbol{p}_i^{\mathrm{T}} & r_i \end{bmatrix}$, then the objective function of (9) becomes $\boldsymbol{N}_i \bullet \boldsymbol{\Gamma}_i$. Restoring the indicator function, the constraint of (9) becomes

$$(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{Q}_i\tilde{\boldsymbol{x}}^i + (\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{p}_i + r_i \geqslant 1,\ \text{if } y^i\left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b\right) \leqslant 1-\xi_i, \tag{10a}$$

$$(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{Q}_i\tilde{\boldsymbol{x}}^i + (\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{p}_i + r_i \geqslant 0,\ \forall\ \tilde{\boldsymbol{x}}^i\in\Xi_i. \tag{10b}$$

Constraint (10b) implies that $[(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\ 1]\boldsymbol{N}_i\begin{bmatrix}\tilde{\boldsymbol{x}}^i\\1\end{bmatrix} \geqslant 0,\ \forall\ \tilde{\boldsymbol{x}}^i\in\Xi_i \Leftrightarrow \boldsymbol{N}_i \succeq 0$. Constraint (10a) can be further transformed as below using the S-Lemma:

$$(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{Q}_i\tilde{\boldsymbol{x}}^i + (\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{p}_i + r_i - 1 + \alpha_i\left(y^i\left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b\right) - 1 + \xi_i\right) \geqslant 0,\ \alpha_i \geqslant 0,$$

which means

$$[(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\ 1]\boldsymbol{N}_i\begin{bmatrix}\tilde{\boldsymbol{x}}^i\\1\end{bmatrix} - 1 + \alpha_i\left(y^i\left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b\right) - 1 + \xi_i\right) \geqslant 0,\ \alpha_i \geqslant 0. \tag{11}$$

If $\alpha_i = 0$, constraint (11) implies that $[(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\ 1]\boldsymbol{N}_i\begin{bmatrix}\tilde{\boldsymbol{x}}^i\\1\end{bmatrix} \geqslant 1$. This contradicts the fact of $\boldsymbol{N}_i \bullet \boldsymbol{\Gamma}_i \leqslant \epsilon, 0<\epsilon<1$, because $\boldsymbol{\Gamma}_i = \mathbb{E}_{F_i}\left[\begin{bmatrix}\tilde{\boldsymbol{x}}^i\\1\end{bmatrix}[(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\ 1]\right]$. Thus, we have $\alpha_i > 0$. Let $\boldsymbol{R}_i = \frac{1}{\alpha_i}\boldsymbol{N}_i$ and $\beta_i = \frac{1}{\alpha_i}$, we see that

$$\begin{aligned}
(11) &\Leftrightarrow [(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\ 1]\frac{\boldsymbol{N}_i}{\alpha_i}\begin{bmatrix}\tilde{\boldsymbol{x}}^i\\1\end{bmatrix} + \left(-\tfrac{1}{\alpha_i} + y^i\left(\tfrac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b\right) - 1 + \xi_i\right) \geqslant 0 \\
&\Leftrightarrow [(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\ 1]\left(\boldsymbol{R}_i + \begin{bmatrix}\frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w}\\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^i b + \xi_i - 1 - \frac{1}{\alpha_i}\end{bmatrix}\right)\begin{bmatrix}\tilde{\boldsymbol{x}}^i\\1\end{bmatrix} \geqslant 0 \\
&\Leftrightarrow \boldsymbol{R}_i + \begin{bmatrix}\frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w}\\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^i b + \xi_i - 1 - \beta_i\end{bmatrix} \succeq 0.
\end{aligned}$$

Therefore, the dual problem (9) becomes

$$\begin{array}{ll}
\inf & \frac{1}{\beta_i}\boldsymbol{\Gamma}_i \bullet \boldsymbol{R}_i \\
s.t. & \boldsymbol{R}_i + \begin{bmatrix}\frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w}\\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^i b + \xi_i - 1 - \beta_i\end{bmatrix} \succeq 0, \\
& \boldsymbol{R}_i \succeq 0, \\
& \boldsymbol{R}_i \in \mathbb{S}^n,\ \beta_i\in\mathbb{R}_+.
\end{array}$$

Since the strong duality holds for the pair of (8) and (9) (Similar to Delage and Ye (2010)), requiring $\varphi_i \leqslant \epsilon$ yields $\frac{1}{\beta_i}\boldsymbol{\Gamma}_i \bullet \boldsymbol{R}_i \leqslant \epsilon \Leftrightarrow \beta_i - \frac{1}{\epsilon}\boldsymbol{\Gamma}_i \bullet \boldsymbol{R}_i \geqslant 0$. This completes the proof. □

Notice that one can reformulate (7) as a standard SDP by rewriting the summation term in the objective function in the matrix form, i.e., $\|\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}\|_2^2 + \frac{C}{N}\xi_i \leqslant \eta_i \Leftrightarrow \begin{bmatrix}\boldsymbol{I}_n & \boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}\\ (\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}} & -\frac{C}{N}\xi_i + \eta_i\end{bmatrix} \succeq 0$, where $\boldsymbol{I}_n$ denotes the n-dimensional identical matrix. This leads to

$$\begin{array}{lll}
\min & \sum_{i=1}^{N}\eta_i & \\
s.t. & \begin{bmatrix}\boldsymbol{I}_n & \boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}\\ (\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}} & -\frac{C}{N}\xi_i + \eta_i\end{bmatrix} \succeq 0, & i=1,\ldots,N,\\
& \boldsymbol{\Gamma}_i \bullet \boldsymbol{R}_i \leqslant \epsilon\beta_i, & i=1,\ldots,N,\\
& \boldsymbol{R}_i + \begin{bmatrix}\frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w}\\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^i b + \xi_i - 1 - \beta_i\end{bmatrix} \succeq 0, & i=1,\ldots,N,\\
& \boldsymbol{R}_i \succeq 0, & i=1,\ldots,N,\\
& \boldsymbol{M}\in\mathbb{S}^n,\ \boldsymbol{w}\in\mathbb{R}^n,\ b\in\mathbb{R},\ \boldsymbol{\xi}\in\mathbb{R}^N_+,\ \boldsymbol{\eta},\ \boldsymbol{\beta}\in\mathbb{R}^N,\ \boldsymbol{R}_i\in\mathbb{S}^{n+1}, & i=1,\ldots,N.
\end{array} \tag{12}$$

In this way, (12) provides an SDP reformulation of (DRC-QSSVM) for using off-the-shelf solvers.

Remark 2.1. Let $\mathcal{V}_i^{SDP}$ denote the feasible region described by the constraints associated with the ith random input in the SDP model, i.e.,

$$\mathcal{V}_i^{SDP} \triangleq \left\{(\boldsymbol{M},\boldsymbol{w},b)\in\mathbb{S}^n\times\mathbb{R}^n\times\mathbb{R} \;\middle|\; \begin{array}{l}\exists\beta_i,\ \boldsymbol{R}_i\succeq 0,\ \beta_i - \frac{1}{\epsilon}\boldsymbol{\Gamma}_i\bullet\boldsymbol{R}_i \geqslant 0,\\ \boldsymbol{R}_i + \begin{bmatrix}\frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w}\\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^i b + \xi_i - 1 - \beta_i\end{bmatrix} \succeq 0\end{array}\right\}. \tag{13}$$

Then Theorem 2.1 implies that $\mathcal{V}_i = \mathcal{V}_i^{SDP}$, for $i=1,\ldots,N$.

2.3. Geometric interpretation

To study the geometric interpretation of the distributionally robust chance constraints in (DRC-QSSVM), we first provide the following lemma:

Lemma 2.2. *For any i, $\mathcal{V}_i^{SDP}$ defined in (13) is equivalent to*

$$\mathcal{V}_i^{SDP'} = \left\{ (\boldsymbol{M},\boldsymbol{w},b)\in\mathbb{S}^n\times\mathbb{R}^n\times\mathbb{R} \;\middle|\; \begin{array}{l} y^i\left(\frac{1}{2}\boldsymbol{d}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{d}_i+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{d}_i+b\right)\geqslant 1-\xi_i-\frac{1}{2}y^i\boldsymbol{M}\bullet\boldsymbol{D}_i^0, \\ \begin{bmatrix}\boldsymbol{\Sigma}_i-\epsilon\boldsymbol{D}_i^0 & \boldsymbol{\mu}_i-\boldsymbol{d}_i\\ \boldsymbol{\mu}_i^{\mathrm{T}}-\boldsymbol{d}_i^{\mathrm{T}} & \frac{1-\epsilon}{\epsilon}\end{bmatrix}\succeq 0,\ \boldsymbol{D}_i^0\in\mathbb{S}_+^n,\ \boldsymbol{d}\in\mathbb{R}^n \end{array} \right\}.$$

Proof. Please see Appendix A.1. □

Now, let $\mathcal{E}(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\frac{1-\epsilon}{\epsilon}) \triangleq \{\boldsymbol{x}\in\mathbb{R}^n \mid (\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\leqslant\frac{1-\epsilon}{\epsilon}\}$ represent an ellipsoid centered at $\boldsymbol{\mu}_i$, whose shape and size are determined by $\boldsymbol{\Sigma}_i$ and ϵ, respectively. Considering the case of correctly classifying $\boldsymbol{x}^i$ for all $\boldsymbol{x}^i\in\mathcal{E}(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\frac{1-\epsilon}{\epsilon})$, we define

$$\mathcal{V}_i^E \triangleq \left\{(\boldsymbol{M},\boldsymbol{w},b)\in\mathbb{S}^n\times\mathbb{R}^n\times\mathbb{R}\middle| y^i\left(\frac{1}{2}(\boldsymbol{x}^i)^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}^i+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i+b\right) \geqslant 1-\xi_i,\forall\,\boldsymbol{x}^i\in\mathcal{E}\left(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\frac{1-\epsilon}{\epsilon}\right)\right\}.$$

In the following lemma, a geometrical interpretation of $\mathcal{V}_i^{SDP}$ is presented by discussing its relation with $\mathcal{V}_i^E$.

Lemma 2.3. *For any given $(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\epsilon)$, $\mathcal{V}_i^E\subseteq\mathcal{V}_i^{SDP}$ for $i=1,\ldots,N$. If $y^i\boldsymbol{M}\succeq 0$, then $\mathcal{V}_i^E=\mathcal{V}_i^{SDP}$. Moreover, for the linear case with $\boldsymbol{M}=\boldsymbol{0}$, $\mathcal{V}_i^E=\mathcal{V}_i^{SDP}$.*

Proof. From Lemma 2.2, the constraints in $\mathcal{V}_i^{SDP}$ can be rewritten as

$$y^i\left(\frac{1}{2}\boldsymbol{d}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{d}_i+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{d}_i+b\right)\geqslant 1-\xi_i-\frac{1}{2}y^i\boldsymbol{M}\bullet\boldsymbol{D}_i^0,\quad \forall\,\boldsymbol{d}^i\in\mathcal{E}\left(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i-\epsilon\boldsymbol{D}_i^0,\frac{1-\epsilon}{\epsilon}\right), \tag{14}$$

with $\boldsymbol{\Sigma}_i-\epsilon\boldsymbol{D}_i^0\succ 0$ and $\boldsymbol{D}_i^0\succeq 0$ being assumed without loss of generality. For each feasible $\boldsymbol{D}_i^0$, this means that the quadratic surface defined by $(\boldsymbol{M},\boldsymbol{w},b)$ separates all points in $\mathcal{E}\left(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i-\epsilon\boldsymbol{D}_i^0,\frac{1-\epsilon}{\epsilon}\right)$ softly. Obviously $\mathcal{V}_i^E$ is a special case of $\mathcal{V}_i^{SDP'}$ when $\boldsymbol{D}_i^0=\boldsymbol{0}$. Therefore, $\mathcal{V}_i^E\subseteq\mathcal{V}_i^{SDP}$. If $y^i\boldsymbol{M}\succeq 0$, we have $y^i\boldsymbol{M}\bullet\boldsymbol{D}_i^0\geqslant 0$ since $\boldsymbol{D}_i^0\succeq 0$. The upper bound of the right-hand side in inequality (14) is obtained when $\boldsymbol{D}_i^0=0$. In this case, we have $\mathcal{V}_i^E=\mathcal{V}_i^{SDP}$. For the linear case with $\boldsymbol{M}=\boldsymbol{0}$, $\boldsymbol{D}_i^0=\boldsymbol{0}$ can be derived from (A.3) in the proof of Lemma 2.2. Thus, $\mathcal{V}_i^E=\mathcal{V}_i^{SDP}$. □

Remark 2.2. Lemma 2.3 implies that the proposed (DRC-QSSVM) model views each uncertain input as a set $\mathcal{E}\left(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i-\epsilon\boldsymbol{D}_i^0,\frac{1-\epsilon}{\epsilon}\right)$ and seeks a maximum-margin classification using these ellipsoids (See Fig. 1). When $\boldsymbol{M}=0$, this result is reduced to the linear case, which leads to a maximum-margin classification using $\mathcal{E}\left(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\frac{1-\epsilon}{\epsilon}\right)$ as discussed by Shivaswamy et al. (2006) and Wang et al. (2018).

Remark 2.3. Note that for each i, the size of the set $\mathcal{E}(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i-\epsilon\boldsymbol{D}_i^0,\frac{1-\epsilon}{\epsilon})$ depends on ϵ. As ϵ decreases, the size increases (See the trends shown by Fig. 1). Consider the following two extreme cases:

- $\epsilon=0$. In this case, $\mathbb{P}_{F_i}\left\{y^i\left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i+\boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i+b\right)\leqslant 1-\xi_i\right\}\leqslant\epsilon=0$, $\forall F_i\in\mathcal{D}_i$, hence, each chance constraint becomes a deterministic constraint, $y^i\left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i+\boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i+b\right)>1-\xi_i$, $\forall F_i\in\mathcal{D}_i$. The quadratic surface $Q(\boldsymbol{M},\boldsymbol{w},b)$ is required to separate data points generated from any potential distribution with fixed mean and covariance. However, there may be numerous possible distributions with the same given mean and covariance, for data points to spread everywhere such that it becomes impossible to find such a separating surface. Also, note that when $\epsilon\to 0$, $\frac{1-\epsilon}{\epsilon}\to\infty$, which means the radius of the ellipsoid implied by $\mathcal{V}_i^{SDP}$ approaches to infinity. It is impossible to find a classifier to separate infinitely large ellipsoids.
- $\epsilon=1$. In this case, $\mathbb{P}_{F_i}\left\{y^i\left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i+\boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i+b\right)\leqslant 1-\xi_i\right\}\leqslant\epsilon=1$, $\forall F_i\in\mathcal{D}_i$, becomes a trivial requirement. This means that for any data point, it is totally random to be correctly classified or not. Then, statistically speaking, the separation surface may be obtained by separating the mean vectors of random data points (See Fig. 1's last column). Note that, $\frac{1-\epsilon}{\epsilon}=0$ as $\epsilon=1$. This result is consistent with Lemma 2.3 when the ellipsoid reduces to the center point $\boldsymbol{\mu}_i$ with a zero radius.

2.4. SOCP reformulation of (DRC-QSSVM)

SDP often requires heavy computational efforts even if it is a tractable convex program, while an equivalent SOCP model may have fewer variables for more efficient computations. This subsection presents SOCP constraints derived from the SDP reformulation of (DRC-QSSVM).

Theorem 2.4. *For $i=1,\ldots,N$, $\mathcal{V}_i^{SDP}$ is equivalent to*

$$\mathcal{V}_i^{SOC} \triangleq \left\{(\boldsymbol{M},\boldsymbol{w},b)\in\mathbb{S}^n\times\mathbb{R}^n\times\mathbb{R}\middle|\begin{array}{l} y^i\left(\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i+\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{w}+b\right)\geqslant 1-\xi_i \\ \quad+\sqrt{\frac{1-\epsilon}{\epsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i+\boldsymbol{w})\|_2-\frac{1-\epsilon}{2\epsilon}\boldsymbol{\Sigma}_i\bullet y^i\boldsymbol{M}\end{array}\right\}. \tag{15}$$

Proof. Please see Appendix A.2. □

Similar to Lemma 2.2, a geometric interpretation of $\mathcal{V}_i^{SOC}$ is given in the next lemma.

Lemma 2.5. *For any given $(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\epsilon)$, $\mathcal{V}_i^E\subseteq\mathcal{V}_i^{SOC}$ for $i=1,\ldots,N$. Moreover, for the linear case when $\boldsymbol{M}=\boldsymbol{0}$, $\mathcal{V}_i^E=\mathcal{V}_i^{SOC}$.*

Proof. Please see Appendix A.3. □

For more efficient computations, we adopt the vectorization technique for $\mathcal{V}_i^{SOC}$. Define the vectorizations of $\boldsymbol{M}\in\mathbb{S}^n$ as $\mathrm{vec}(\boldsymbol{M})\triangleq[M_{11},\ldots,M_{1n},M_{21},\ldots,M_{2n},M_{n1},\ldots,M_{nn}]^{\mathrm{T}}\in\mathbb{R}^{n^2}$, and $\mathrm{hvec}(\boldsymbol{M})\triangleq[M_{11},\ldots,M_{1n},M_{22},\ldots,M_{2n},M_{n-1,n-1},M_{n-1,n},M_{nn}]^{\mathrm{T}}\in\mathbb{R}^{\frac{n(n+1)}{2}}$. Let $\boldsymbol{D}_n\in\mathbb{R}^{n^2\times\frac{n(n+1)}{2}}$ be the matrix that satisfies $\boldsymbol{D}_n\mathrm{hvec}(\boldsymbol{M})=\mathrm{vec}(\boldsymbol{M})$. For any i, define

$$\boldsymbol{H}^i\triangleq[\boldsymbol{I}_n\otimes(\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{D}_n \quad \boldsymbol{I}_n]\in\mathbb{R}^{n\times(\frac{n(n+1)}{2}+n)}, \tag{16}$$

where $\otimes$ denotes the Kronecker product.

Let $\boldsymbol{z}=[\mathrm{hvec}(\boldsymbol{M})^{\mathrm{T}}\ \boldsymbol{w}^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{\frac{n(n+1)}{2}+n}$ be the reorganized variable of $\boldsymbol{M}$ and $\boldsymbol{w}$. The objective function of (7) can be rewritten as $\sum_{i=1}^N\|\boldsymbol{M}\boldsymbol{\mu}_i+\boldsymbol{w}\|_2^2=\sum_{i=1}^N(\boldsymbol{H}^i\boldsymbol{z})^{\mathrm{T}}(\boldsymbol{H}^i\boldsymbol{z})=\boldsymbol{z}^{\mathrm{T}}(\sum_{i=1}^N(\boldsymbol{H}^i)^{\mathrm{T}}\boldsymbol{H}^i)\boldsymbol{z}$. Letting $\boldsymbol{W}=\sum_{i=1}^N(\boldsymbol{H}^i)^{\mathrm{T}}\boldsymbol{H}^i$, we further have $\sum_{i=1}^N\|\boldsymbol{M}\boldsymbol{\mu}_i+\boldsymbol{w}\|_2^2=\boldsymbol{z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{z}$. Let $\boldsymbol{V}_{\frac{n(n+1)}{2}}\triangleq 2\boldsymbol{I}_{\frac{n(n+1)}{2}}-\mathrm{Diag}(\mathrm{hvec}(\boldsymbol{I}_n))$. Then we define

$$\boldsymbol{r}^i\triangleq[\frac{1}{2}(\boldsymbol{V}_{\frac{n(n+1)}{2}}\mathrm{hvec}(\frac{1-\epsilon}{\epsilon}\boldsymbol{\Sigma}_i+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}))^{\mathrm{T}}\ \boldsymbol{\mu}_i^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{\frac{n(n+1)}{2}+n}. \tag{17}$$

We have $\frac{1}{2}\boldsymbol{M}\bullet(\frac{1-\epsilon}{\epsilon}\boldsymbol{\Sigma}_i+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}})+\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{w}=\boldsymbol{z}^{\mathrm{T}}\boldsymbol{r}^i$. Hence, (DRC-QSSVM) with a feasible set in the form of $\mathcal{V}_i^{SOC}$ has the following SOCP reformulation:

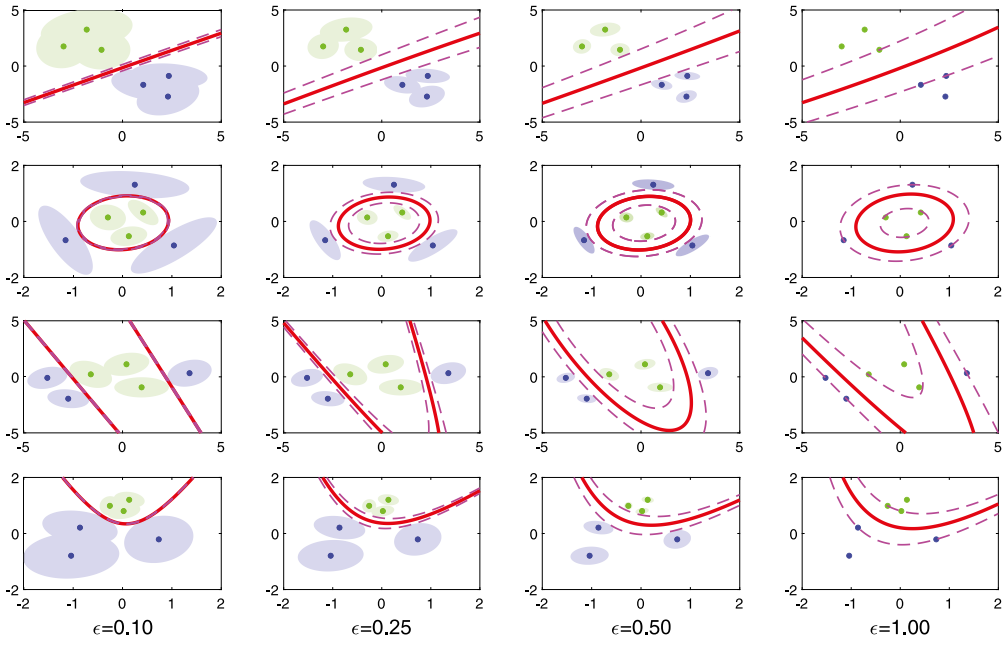


Fig. 1. Geometric interpretation of (DRC-QSSVM) on 2D synthetic data. The green and blue solid points represent $\{\boldsymbol{\mu}_i\}_{i=1}^N$ in two classes, respectively. Shaded areas depict the corresponding ellipsoids $E(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \frac{1-\epsilon}{\epsilon})$. The learned quadratic classifier, $\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}+b=0$, is represented by the red solid line, and the pink dashed lines represent quadratic curves defined by $\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}+b=\pm 1$.

$$
\begin{aligned}
\min\quad & \boldsymbol{z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{z}+C\sum_{i=1}^{N}\xi_i \\
s.t.\quad & y^i\left(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{r}^i+b\right)\geqslant 1-\xi_i+\sqrt{\tfrac{1-\epsilon}{\epsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{H}^i\boldsymbol{z}\|_2,\ i=1,\ldots,N, \\
& \boldsymbol{z}\in\mathbb{R}^{\frac{n(n+1)}{2}+n},\ b\in\mathbb{R},\ \boldsymbol{\xi}\in\mathbb{R}_+^N.
\end{aligned}
\tag{18}
$$

Since matrix $\boldsymbol{W}$ is real, symmetric, and positive semi-definite, its Cholesky factorization leads to $\boldsymbol{W}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}$ with $\boldsymbol{Q}\in\mathbb{S}_+^{\frac{n(n+1)}{2}+n}$. Consequently, we have the following standard SOCP formulation:

$$
\begin{aligned}
\min\quad & \theta+C\sum_{i=1}^{N}\xi_i \\
s.t.\quad & \|\boldsymbol{Q}\boldsymbol{z}\|_2\leqslant\theta, \\
& \|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{H}^i\boldsymbol{z}\|_2\leqslant\sqrt{\tfrac{\epsilon}{1-\epsilon}}\left(y^i\left(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{r}^i+b\right)-1+\xi_i\right),\ i=1,\ldots,N, \\
& \boldsymbol{z}\in\mathbb{R}^{\frac{n(n+1)}{2}+n},\ b\in\mathbb{R},\ \theta\in\mathbb{R},\ \boldsymbol{\xi}\in\mathbb{R}_+^N.
\end{aligned}
\tag{19}
$$

Some observations on the SDP model (12) and SOCP model (19) can be made here.

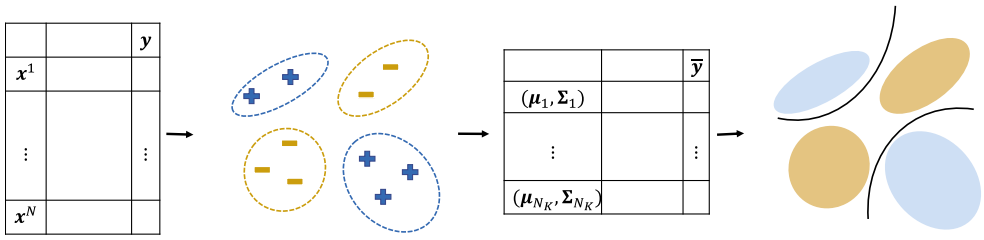
- Excluding the common variables $(b,\boldsymbol{\xi})\in\mathbb{R}\times\mathbb{R}_+^N$, the SDP model has N matrix variables in $\mathbb{S}^{n+1}$, one matrix variable in $\mathbb{S}^n$, two vector variables in $\mathbb{R}^N$ and one vector variable in $\mathbb{R}^n$, while the SOCP model only has one vector variable in $\mathbb{R}^{\frac{n(n+1)}{2}+n}$ and one scalar variable.
- The SDP model has $3N$ semidefinite constraints involving $(n+1)$-dimensional positive semi-definite cones, while the SOCP model only has one constraint involving $(n(n+1)/2+n+1)$-dimensional second-order cones and N constraints involving $(n+1)$-dimensional second-order cones.

For general practice, the number of features n is much smaller than that of input points N. Therefore, the SOCP model is more computationally friendly than the SDP model.

3. Data-driven approach for data sets without moment information

The results presented in the previous section count on the first- and second-order moments of the data set with probability uncertainty. This section utilizes the proposed (DRC-QSSVM) for classifying data sets without moment information. Fig. 2 illustrates the basic workflow of our data-driven approach. For a finite set of binary samples $\{(\boldsymbol{x}^i,y^i)\in\mathbb{R}^n\times\{-1,1\},\ i=1,\ldots,N\}$ without any moment information (Fig. 2a), we intend to use the proposed (DRC-QSSVM) model for building a robust classifier. The key idea is to group similar data points together and represent them by the mean and covariance of the group. Clustering is widely used to identify the inherent structure of data, while it can also serve as a pre-processing technique for classification (Zhou, 2021). Clustering aims to partition a data set into disjoint subsets, called clusters, where data points within the same cluster have high similarities. As shown in Fig. 2(b), we first use clustering algorithms to partition the given data set into N_K clusters, $\boldsymbol{C}_k$, $k=1,\ldots,N_K$, with $N_K<N$. The K-means clustering algorithm could partition data into a finite number of homogeneous and separate clusters without using any prior knowledge. Hence we adopt the K-means++ algorithm proposed by Vassilvitskii and Arthur (2006) that enhances the performance of ordinary K-means algorithms. The well-known ELBOW validation method could help determine an optimal value of N_K. Once the clusters were obtained, the sample mean and covariance matrix can be easily calculated to estimate $(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$, $k=1,\ldots,N_K$ (Fig. 2(c)). Moreover, the sample mean is denoted as $\bar{\boldsymbol{x}}^k\triangleq\frac{1}{|\boldsymbol{C}_k|}\sum_{i=1}^N\boldsymbol{x}^i\mathbb{1}_{\boldsymbol{C}_k}(\boldsymbol{x}^i)$, and the sample covariance as $\boldsymbol{S}^k\triangleq\frac{1}{|\boldsymbol{C}_k|-1}\sum_{i=1}^N(\boldsymbol{x}^i\mathbb{1}_{\boldsymbol{C}_k}(\boldsymbol{x}^i)-\bar{\boldsymbol{x}}^k)(\boldsymbol{x}^i\mathbb{1}_{\boldsymbol{C}_k}(\boldsymbol{x}^i)-\bar{\boldsymbol{x}}^k)^{\mathrm{T}}$, where $|\boldsymbol{C}_k|$ is the cardinality of the kth cluster $\boldsymbol{C}_k$. Then we could construct the ambiguity set (3) to employ the proposed (DRC-QSSVM) model for a robust classifier described in Fig. 2(d).

The proposed process illustrated by Fig. 2 leads to Algorithm 1, where Step 1 partitions the given binary data shown in 2(a) into two classes first. The clustering shown in Fig. 2(b) is accomplished by Steps 2–8. Step 9 computes the mean and covariance shown in Fig. 2(c). Steps 10–12 employ the proposed model to obtain a robust quadratic



(a) Dataset (b) Clustering (c) 1st- and 2nd-Moments (d) Classifier

Fig. 2. Data-driven approach for data sets without moment information.

classifier (Fig. 2(d)) determined by $(\boldsymbol{M}^*, \boldsymbol{w}^*, b^*)$ by solving the SOCP reformulation (18).

Algorithm 1 Data-driven distributionally robust classification algorithm.

Input: Data set $\boldsymbol{C} = \{(\mathbf{x}^i, y^i) \in \mathbb{R}^n \times \{-1, 1\}, i = 1, \ldots, N\}$, $\epsilon \in (0, 1)$.
Output: A quadratic surface defined by $(\boldsymbol{M}^*, \boldsymbol{w}^*, b^*)$.

1: Extract two classes from the original data set $\boldsymbol{C}$:
$\boldsymbol{C}_+ = \{\boldsymbol{x}^i \in \mathbb{R}^n \mid y^i = 1, \quad i = 1, \ldots, N\}$. $\boldsymbol{C}_- = \{\boldsymbol{x}^i \in \mathbb{R}^n \mid y^i = -1, i = 1, \ldots, N\}$.
2: Compute the cardinality: $N_+ = |\boldsymbol{C}_+|$, $N_- = |\boldsymbol{C}_-|$.
3: Let N_+^U be the upper bound of possible cluster numbers for $\boldsymbol{C}_+$, and N_-^U for $\boldsymbol{C}_-$.
Set $N_+^U = \min\{25, 10\% N_+\}$, and $N_-^U = \min\{25, 10\% N_-\}$.
The values 25 and 10% are user-defined based on the size of the data set.
4: **for** $I \in \{$ '+', '−'$\}$ **do**
5: **for** $K = 1, 2, \ldots, N_I^U$ **do**
6: Apply the K-means++ algorithm to divide the set $\boldsymbol{C}_I$ into K clusters.
7: **end for**
8: Use the Elbow method to find the optimal value of K denoted as K_I^*.
9: **end for**
10: Set $N_K = K_+^* + K_-^*$.
11: Calculate $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ for $\boldsymbol{C}_+^k$ and number them by $k = 1, \ldots, K_+^*$. Calculate $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ for $\boldsymbol{C}_-^k$ and number them by $k = K_+^* + 1, \ldots, N_K$. Set $\bar{\boldsymbol{y}} = [\boldsymbol{e}_{K_+^*}; -\boldsymbol{e}_{K_-^*}] \in \mathbb{R}^{N_K}$ as the new label vector, where $\boldsymbol{e}_{K_+^*}$ and $\boldsymbol{e}_{K_-^*}$ are K_+^*-dimensional and K_-^*-dimensional vectors of all ones, respectively.
12: Use $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ to compute $\boldsymbol{H}^k$ by (16), $\boldsymbol{r}^k$ by (17), $k = 1, \ldots, N_K$. Set $\boldsymbol{W} = \sum_{k=1}^{N_K} (\boldsymbol{H}^k)^{\mathrm{T}} \boldsymbol{H}^k$.
13: With $\boldsymbol{W}$, $\boldsymbol{H}^k$, $\boldsymbol{r}^k$, $k = 1, \ldots, N_K$, and $\bar{\boldsymbol{y}}$, solve the SOCP problem (18) of (DRC-QSSVM) to find an optimal solution $(\boldsymbol{z}^*, b^*)$.
14: Compute $(\boldsymbol{M}^*, \boldsymbol{w}^*)$ from $\boldsymbol{z}^*$ using the vectorization technique discussed in Section 2.4.
15: **Return** $(\boldsymbol{M}^*, \boldsymbol{w}^*, b^*)$.

Algorithm 1 implies that the robust quadratic classifier is learned by separating N_K ellipsoids instead of separating N data points with $N_K \ll N$ in general. This observation indicates the potential benefit of our proposed in dealing with massive data that shall be explored by computational experiments in Section 4.2.2. In addition, we notice that the classification objects become ellipsoids that cover data points with similarity, which might help avoid outliers and balance the magnitude of the two classes. It implies the potential of Algorithm 1 for treating imbalanced data. This works especially well for applications on rare case detection. SVMs with non-robust counterparts may perform poorly on the minority class since they may focus on the majority class while maximizing the overall accuracy (Thabtah, Hammoud, Kamalov, & Gonsalves, 2020). To illustrate this idea, for a highly imbalanced data set with a ratio of 5/21, Fig. 3(a) shows that the quadratic classifier obtained by (QSSVM) sacrifices 4 minority points in brown by treating them as outliers. However, instead of classifying the 26 imbalanced points, the proposed approach finds a quadratic classifier by separating 4 ellipsoids with a more balanced ratio of 2/2 (See Fig. 3(b)). Further experiments on classifying imbalanced data will be conducted in Section 4.2.3.

Remark 3.1. In practice, decision-makers can rarely be completely confident in the sample mean and covariance matrix for estimated moments. The challenge here is that the sample means and covariance matrices themselves are uncertain. Hence their uncertainty is factored into chance constraints by considering a confidence region for the sample mean and covariance matrix. Appendix B.1 further extends the work.

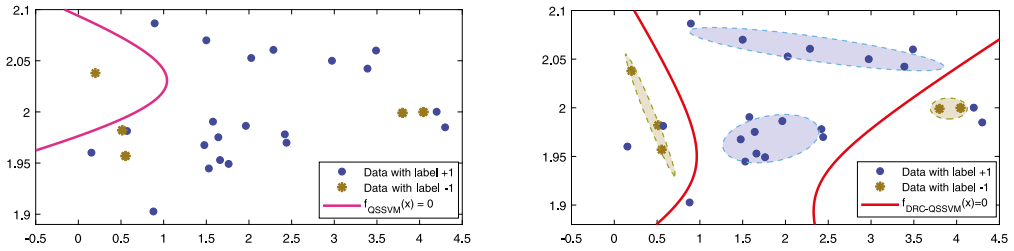
Remark 3.2. As shown in Fig. 2 and Algorithm 1, the data-driven approach first takes the clustering process and then extracts the moment information. For real-world applications, historical data might not be sufficient to reflect the whole data structure, and we may need to collect new data over time to form a dynamic approach. When new instances are added to the data set, we need to update the clustering and classification processes as well.

4. Computational experiments

This section studies the proposed (DRC-QSSVM) model by computational experiments. For uncertain data with given first- and second-order moments, in Section 4.1, we validate the effectiveness of the proposed model and analyze its performance in terms of the parameter ϵ. Then we compare the proposed model with the DRC linear soft SVM (DRC-LSSVM) model which is the only maximum-margin SVM model using the first- and second-order moments information in the literature. For data without moment information, in Section 4.2, we compare the data-driven approach proposed in Section 3 with some state-of-the-art SVM models. In particular, we explore the potential benefits of using the proposed model for problems with massive and/or imbalanced data. In this section, all computational experiments were conducted using MATLAB (R2021a) software on a desktop equipped with Intel(R) Core(TM) i3-9100 CPU @ 3.60 GHz CPUs and 32 GB RAM. The commercial solver SDPT3 (Toh, Todd, & Tütüncü, 1999) is employed to solve SDP and SOCP problems.

4.1. Data sets with first- and second-order moments

As discussed in Section 2, for uncertain data with first- and second-order moments, the proposed (DRC-QSSVM) model has computable SOCP and SDP reformulations. In Section 4.1.1, we test these two formulations on synthetic data sets regarding classification accuracy and computational efficiency. For the proposed model, the parameter C controls the trade-off between maximizing the margin and minimizing the misclassification loss, as commonly adopted in most SVM models. While the parameter ϵ determines the upper bound of misclassification probability that affects the quality of robust classification. Here we skip the detailed analysis on the parameter C, but focus on the parameter



(a) Result by QSSVM. (b) Result by DRC-QSSVM.

Fig. 3. Comparison on an imbalanced data set.

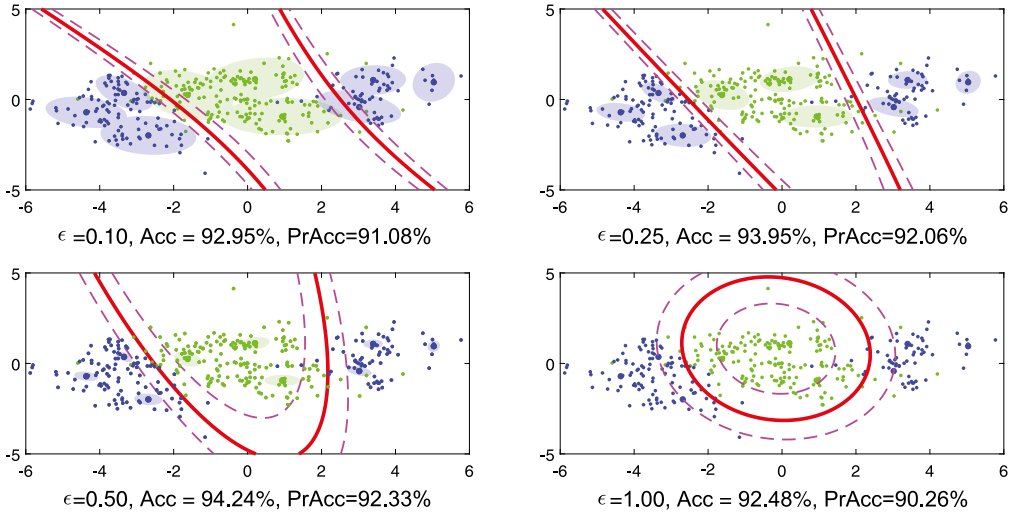


Fig. 4. Results of (DRC-QSSVM) on Syn-Hype-2d-9 data. Shaded areas depict the given ellipsoids $E(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \frac{1-\epsilon}{\epsilon})$. The learned quadratic classifier, $\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}+b=0$, is represented by the red solid line, and the pink dashed lines represent $\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}+b=\pm 1$. Solid points represent random testing points.

ϵ. For all computational experiments, we take the grid method to set $C \in \{2^{-1}, 2^{1}, \ldots, 2^{14}\}$ and $\epsilon \in \{0.10, 0.25, 0.50, 1.00\}$.

We first introduce some error measures to be used. For a quadratic surface obtained by solving (DRC-QSSVM) with an output $(\boldsymbol{M}^*, \boldsymbol{w}^*, b^*) \in \mathbb{S}^n \times \mathbb{R}^n \times \mathbb{R}$, the predicted label $\hat{y} = \mathrm{sign}(\frac{1}{2}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{M}^*\boldsymbol{x} + (\boldsymbol{w}^*)^{\mathrm{T}}\boldsymbol{x} + b^*)$ can be determined for a test data point $(\boldsymbol{x}, y)$. A commonly used measure is the accuracy score (Acc) computed by $\sum_{i=1}^{N} \mathbb{1}(\hat{y}^i = y^i)/N \times 100\%$ where N is the total number of tested points. However, for the uncertain classification with data points from a distribution, we further consider a probabilistic accuracy score (PrAcc). Note that Ben-Tal et al. (2011) and Wang et al. (2018) adopted an "optimal error" to quantify the probabilistic error. For DRC-LSSVM, we have PrAcc = 1− "optimal error". Here we further extend the measure for quadratic classifiers. At each uncertain data point $\tilde{\boldsymbol{x}}^i$ associated with the ambiguity set D_i in (3), the robust chance constraint, $\sup_{F_i \in D_i} \mathbb{P}_{F_i}\left\{ y^i \left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i + \boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i + b \right) \leqslant 0 \right\} \leqslant \epsilon$, ensures an upper bound ϵ of the misclassification probability for the quadratic classifier. Using Theorem 2.4 and the SOCP reformulation (19), the true probability of misclassification at $\tilde{\boldsymbol{x}}^i$ should be no more than ϵ, if $\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{T}^i\boldsymbol{z}\|_2 \leqslant \sqrt{\epsilon/(1-\epsilon)}y^i(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{r}^i + b)$. And this could imply that the least value of ϵ for $\tilde{\boldsymbol{x}}^i$ is $\epsilon_i^* = \frac{\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{T}^i)^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{T}^i\boldsymbol{z}}{(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{r}^i+b)^2+\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{T}^i)^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{T}^i\boldsymbol{z}}$. Consequently, we define

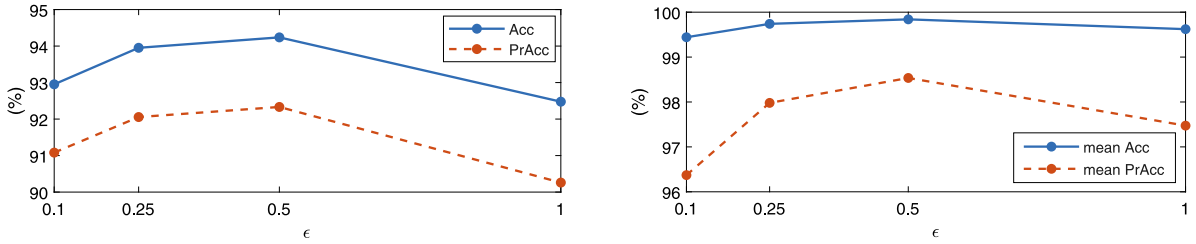
$$\mathrm{PrAcc} = (1 - \sum_{i=1}^{N} \mathrm{err}_i/N) \times 100\%, \quad \text{where } \mathrm{err}_i = \begin{cases} 1, & \text{if } \hat{y}^i \neq y^i \\ \epsilon_i^*, & \text{if } \hat{y}^i = y^i \end{cases}, i = 1, \ldots, N. \tag{20}$$

4.1.1. Validation and analysis of the proposed model

Given that we have uncertain data set with known $\{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \in \mathbb{R}^n \times \mathbb{S}^n, i = 1, \ldots, N\}$, to validate the proposed model and its effectiveness, we first generate some synthetic data in different quadratic patterns including hyperbolic, elliptic, and parabolic structures (See Fig. 1 for illustration). The corresponding synthetic data sets are named in the format of "Syn-Pattern-nd-N", for example, the data set "Syn-Hype-4d-50" denotes a set of $\{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \in \mathbb{R}^4 \times \mathbb{S}^4, i = 1, \ldots, 50\}$ where $\boldsymbol{\mu}_i$ are generated along with a hyperbolic surface, and $\boldsymbol{\Sigma}_i$ are random positive definite matrices with eigenvalues in $[0, 1]$. For each i, we generate 50 random points following the normal distribution with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$ as the testing data points. Note that the proposed model can handle distribution-free data and we choose the normal distribution for simplicity in this section. We generate 8 data sets by selecting $n \in \{2, 4, 8, 16\}$ and $N \in \{50, 100\}$. We also record the testing accuracy by the average Acc and PrAcc. The average training CPU time is recorded for both results solved by the SDP model (12) and SOCP model (19).

For a simple illustration, first, we show a 2-dimensional example tested on the "Syn-Hype-2d-9" data set. Fig. 4 shows that the classifiers learned based on the first- and second-order moments depend on the value of ϵ. For example, the proposed model provides hyperbolic curves when $\epsilon = 0.10, 0.25$, a parabolic curve when $\epsilon = 0.50$, and an ellipsoidal curve when $\epsilon = 1.00$. Figs. 4 and 5(a) show that ϵ indeed affects the classification accuracy. The Acc and PrAcc shown in Fig. 5(a) depict how ϵ affects the performance.

Synthetic data sets with bigger sample sizes in higher dimensions have been tested to investigate the proposed model further. Table 1 shows one group of results on "hyperbolic" synthetic data sets. More results on the "elliptic" and "parabolic" data sets are shown in Appendix B.2. For all testing problems, we ensure that the SOCP model



(a) Syn-Hype-2d-9 data. (b) Data tested in Table 1.

Fig. 5. Performance of (DRC-QSSVM) on "hyperbolic" synthetic data sets in terms of ϵ.

Table 1
Testing results on "hyperbolic" synthetic data sets by DRC-QSSVM.

Data		Syn-Hype-4d-50				Syn-Hype-8d-50				Syn-Hype-16d-50			
ϵ		0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
Acc(%)		99.96	100.00	99.96	99.24	97.92	98.64	99.28	99.04	98.88	99.84	99.92	99.68
PrAcc(%)		98.54	99.18	99.39	98.69	92.57	95.49	96.46	93.61	88.67	94.76	96.94	94.11
CPU time (s)	SOCP	3.01	2.99	2.99	2.22	3.14	3.13	3.08	2.38	5.72	5.71	5.98	3.85
	SDP	4.55	4.50	4.45	3.35	6.98	6.99	6.89	5.10	34.96	30.90	29.54	21.77

Data		Syn-Hype-4d-100				Syn-Hype-8d-100				Syn-Hype-16d-100			
ϵ		0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
Acc(%)		100.00	99.96	99.88	99.88	99.89	100.00	100.00	100.00	100.00	100.00	100.00	99.89
PrAcc(%)		99.35	99.36	99.27	99.26	99.59	99.70	99.80	99.83	99.51	99.39	99.35	99.34
CPU time (s)	SOCP	5.14	5.25	5.22	2.97	5.50	5.56	5.62	3.89	12.43	12.41	12.36	7.70
	SDP	8.05	8.03	7.95	4.74	19.40	18.20	17.81	11.74	187.10	176.16	142.67	136.87

achieves the same results as the SDP model, but in a much more efficient way. From Table 1, we see that (i) the proposed (DRC-QSSVM) model performs well in terms of Acc and PrAcc measures; (ii) it is not surprising that PrAcc is always smaller than the corresponding Acc since the former considers the potential misclassification probability when the predicted label is correct; (iii) For fixed N and n, both Acc and PrAcc change depending on ϵ, but in the same trend, as illustrated in Fig. 5(b).

4.1.2. Comparison with the DRC-LSSVM model

In the literature, DRC-LSSVM (Wang et al., 2018) is the only maximum-margin SVM model using the means and covariance matrices for distributionally robust classification. Hence we compare the proposed (DRC-QSSVM) with DRC-LSSVM using the well-known data sets Wisconsin breast cancer (WIBC) and the Ionosphere from the UCI dataset. For fair comparisons, we adopt the same settings for data preprocessing as in Wang et al. (2018). WIBC data contains 683 samples with 9 features, i.e. $N = 683$, $n = 9$, and extracted Ionosphere data has $N = 351$, $n = 15$ (extracted from $n = 34$ as Wang et al. (2018)). Moreover, for computational efficiency, SOCP reformulations are used for both (DRC-QSSVM) and DRC-LSSVM. Similarly to Wang et al. (2018), $\boldsymbol{\mu}_i$ is set to be the value of each training point, and $\boldsymbol{\Sigma}^i$ is calculated based on the covariance matrix of all training points in the same class. Table 2 shows the results where (i) 20% of the data are used for training and the remaining 80% for testing; (ii) 80% for training and 20% for testing.

The proposed DRC-QSSVM model generates a quadratic surface, which increases the flexibility when handling nonlinear data. However, it increases the model complexity as well compared with the linear DRC-LSSVM model. Table 2 clearly shows that the performance of (DRC-QSSVM) dominates that of DRC-LSSVM in all cases, taking a reasonably longer running time. The superiority of the proposed (DRC-QSSVM) model becomes particularly evident when applied to the Extracted Ionosphere data, which is more nonlinearly complex than the WIBC data.

4.2. Data sets without moment information

Section 3 provides a data-driven approach to apply (DRC-QSSVM) for robustly classifying exact data points $\{(\boldsymbol{x}^i, y^i) \in \mathbb{R}^n \times \{-1, +1\}, i = 1, \ldots, N\}$. We conduct computational experiments to compare the data-driven approach with well-known state-of-the-art SVMs using some commonly used public benchmark data sets. Table 3 lists the tested models, including their abbreviations, solvers, and parameters. Kernelized SVMs are solved by utilizing LIBSVM (Chang & Lin, 2011), and other SVMs are solved by SDPT3. Note that (DRC-QSSVM) is realized by the data-driven Algorithm 1 in which the SOCP problem (18) is solved by SDPT3.

For all tests, the 10-fold cross-validation and grid methods are adopted to select the best parameters of C, ϵ, and σ from the ranges of $C \in \{2^{-1}, 2^1, \ldots, 2^{14}\}$, $\epsilon \in \{0.1, 0.2, \ldots, 1\}$, and $\sigma \in \{2^{-5}, 2^{-4}, \ldots, 2^5\}$, respectively. All test results are based on the best-selected parameters. Some public benchmark data sets from UCI databases (See Table 4) are chosen. Section 4.2.1 presents the results of "balanced" data sets, including the commonly used Scale, Pima Indians Diabetes (Pima), WIBC, and Ionosphere. Section 4.2.2 reports the performance of "massive" data sets including Skin and Cod-RNA with large sample sizes. Section 4.2.3 explores the results of "imbalanced" data sets, including Car Evaluation (Careval) and Heart Disease (Heart) with skewed class proportions. All the classical SVMs are tested on the original data, and the two robust models including DRC-LSSVM and the proposed DRC-QSSVM utilize the moment information of the data retrieved by the data-driven approach described in Algorithm 1.

4.2.1. Benchmark tests on balanced data

Four popular balanced benchmark data sets: Scale, Pima, WIBC, and Ionosphere are tested. Since they are exact data sets, we report the Acc

Table 2
Testing results on WIBC and extracted Ionosphere data sets.

Data (20% training)		WIBC				Extracted Ionosphere			
ϵ		0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
DRC-LSSVM	Acc(%)	96.60	96.38	96.30	96.12	84.26	84.54	84.40	84.26
	PrAcc(%)	93.98	93.70	93.61	93.00	82.61	82.91	82.09	81.92
	CPU time (s)	5.77	5.48	5.05	5.01	3.36	3.37	3.39	3.40
DRC-QSSVM	Acc(%)	96.63	96.63	96.52	96.56	92.81	91.95	92.09	92.09
	PrAcc(%)	94.05	94.39	94.69	94.82	91.64	91.21	91.38	91.30
	CPU time (s)	6.66	6.13	5.71	5.77	5.56	5.53	5.56	5.61
Data (80% training)		WIBC				Extracted Ionosphere			
ϵ		0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
DRC-LSSVM	Acc(%)	96.63	96.63	96.49	96.49	87.46	88.31	88.02	87.74
	PrAcc(%)	95.45	95.57	95.35	95.15	83.36	83.70	84.19	83.67
	CPU time (s)	16.97	17.56	16.37	18.14	10.51	10.54	12.15	11.37
DRC-QSSVM	Acc(%)	97.22	97.37	97.22	96.93	96.58	96.02	95.73	95.16
	PrAcc(%)	95.86	95.91	95.76	95.61	92.87	93.48	92.94	92.18
	CPU time (s)	32.25	30.82	34.77	30.51	33.05	32.77	30.47	31.01

Table 3
Models and solvers of the tested models.

Model	Abbreviation	Solver/Package	Parameter
Linear soft SVM	LSSVM	LibSVM	C
Quadratic soft surface SVM	QSSVM	SDPT3	C
SVM with quadratic kernel	KQSSVM	LibSVM	(C,σ)
SVM with Gaussian kernel	KGSSVM	LibSVM	(C,σ)
DRC linear SVM	DRC-LSSVM	SDPT3	(C,ϵ)
The proposed model	DRC-QSSVM	SDPT3	(C,ϵ)

measure only. The mean and standard deviation of Acc are shown in Table 5. Same as in Section 4.1.2, we select 20% and 80% of data sets for training, respectively. The average training CPU time of each model is also reported.

The following observations can be made:

- The proposed (DRC-QSSVM) model produces much more accurate classifications than other tested SVM models on all tested balanced data sets. It shows the special value of the data-driven based robust (DRC-QSSVM) for commonly used data sets without prescribed uncertainty.
- For most data sets, the classification accuracy obtained by (DRC-QSSVM) changes very little in terms of different training rates of 20% vs 80%. It indicates the potential advantage of the proposed model when we have limited data points for training.
- The CPU time consumed by the proposed (DRC-QSSVM) is acceptable overall considering its classification accuracy. Also note that (DRC-QSSVM), (QSSVM), and DRC-LSSVM are solved using the solver SDPT 3.0, while others are solved using an integrated software LIBSVM.

4.2.2. Benchmark tests on massive data

Two massive benchmark data sets, Skin and Cod-RNA, are used for testing. We also use 20% and 80% of data points for training, respectively. The mean, standard deviation of accuracy scores, and the average training CPU time are reported in Table 6.

The following observations can be made:

- The proposed (DRC-QSSVM) model significantly outperforms others in accuracy for the Skin data. For the Cod-RNA data, (DRC-QSSVM) also outperforms other models considering both the accuracy and CPU time.
- (DRC-QSSVM) can achieve high accuracy using only 20% data points for training. This supports the stability of the proposed model and the promising potential of practical use for massive data classification.

4.2.3. Benchmark tests on imbalanced data

Imbalanced data sets, where one class greatly outnumbers the other, are a common issue in many real-life applications. Classifying imbalanced data presents a challenge for standard classification algorithms. In this subsection, two imbalanced data sets, Careval and Heart, are tested. For the classification of imbalanced data, a good SVM model should (i) enhance recognition success specifically for the minority class, and/or (ii) balance recognition capabilities between both classes (Sun, Wong, & Kamel, 2009). Additional error measurements are often used to evaluate such performance. The Area Under Curve (AUC) score could help evaluate the first performance, while the G-mean score could help the second one (Details refer to Sun et al. (2009)). In this subsection, we elect 80% as the training rate due to the potential inadequacy of the minority class sample size to facilitate training at the 20% rate. Table 7 displays the average CPU time, and the mean and standard deviation of the Acc, AUC, and G-mean scores.

The following observations can be made:

- The proposed (DRC-QSSVM) model outperforms other models in all three measures. The dominance is particularly significant in AUC and G-mean scores.
- Note that the imbalance ratio of the Heart data is higher than that of the Careval data. Most models have AUC and G-mean around 50%, which means these classifiers cannot distinguish two classes clearly. However, the proposed model still shows good performance. This means that the proposed model may have a better capability of handling highly imbalanced data.

In summary, applying the DRO approach, the proposed model extends the QSSVM framework, enabling a kernel-free nonlinear SVM with a quadratic classifier to handle uncertainties in data. It outperforms the QSSVM as well as other state-of-the-art SVMs in general classification tasks without uncertainty. While the QSSVM is tested on original data, solving a certain problem, the proposed model leverages hidden moment information to address uncertain problems. This highlights the significant finding that transforming a certain problem into an uncertain one and then solving it may lead to surprisingly better outcomes.

Table 4
Summary of the benchmark data sets.

Data set		Balanced				Massive		Imbalanced	
		Scale	Pima	WIBC	Ionosphere	Skin	Cod-RNA	Careval	Heart
Dimension	n	4	8	9	34	3	8	6	9
Sample size[a] N	N_+	288	268	239	225	50,859	19,845	1,210	3,101
	N_-	288	500	444	126	194,198	39,690	384	557

[a] $N_\pm$ = sample size of points in the class labeled '±1' and $N = N_+ + N_-$.

Table 5
Testing results on the balanced benchmark data sets.

Model/Training rate \ Data set		Scale			Pima			WIBC			Ionosphere		
		Acc(%)		CPU(s)	Acc(%)		CPU(s)	Acc(%)		CPU(s)	Acc(%)		CPU(s)
		mean	std		mean	std		mean	std		mean	std	
LSSVM	20%	70.24	6.99	0.27	74.97	3.69	0.01	92.65	3.16	0.01	82.48	1.44	0.01
	80%	73.02	5.63	1.50	76.14	2.51	0.08	96.03	1.26	3.90	88.76	3.19	0.28
QSSVM	20%	97.35	0.89	0.71	74.31	3.87	0.78	94.12	2.38	2.44	87.81	2.64	3.18
	80%	97.53	0.91	8.50	76.67	2.43	9.87	95.66	0.95	11.58	93.52	2.97	10.29
KQSSVM	20%	97.65	0.78	0.16	76.99	2.68	0.23	95.44	1.66	0.15	87.43	3.12	0.05
	80%	97.65	0.88	1.11	78.24	1.97	1.01	96.10	0.98	0.54	91.62	4.06	0.13
KGSSVM	20%	97.82	1.00	0.01	76.34	2.09	0.01	92.18	1.42	0.01	88.57	1.56	0.01
	80%	98.24	0.83	0.02	77.58	2.09	0.11	95.96	0.93	0.02	93.74	1.17	0.02
DRC-LSSVM	20%	68.65	5.03	0.40	76.99	2.80	0.72	94.49	2.00	2.10	88.19	3.73	0.50
	80%	74.18	4.59	0.48	76.93	3.28	0.77	94.71	1.54	2.03	89.05	5.33	0.74
DRC-QSSVM	20%	**98.18**	1.01	0.44	**80.77**	4.15	0.77	**96.55**	2.03	2.20	**93.52**	1.17	3.04
	80%	**98.30**	0.79	0.51	**81.36**	3.42	0.82	**96.66**	1.29	2.14	**94.00**	2.97	3.57

Table 6
Testing results on the massive benchmark data sets.

Training rate	Skin						Cod-RNA					
	20%			80%			20%			80%		
	Acc (%)		CPU(s)	Acc (%)		CPU(s)	Acc (%)		CPU(s)	Acc (%)		CPU(s)
	mean	std		mean	std		mean	std		mean	std	
LSSVM	75.01	8.31	1.52	80.01	5.29	6.08	78.51	10.32	10.88	84.67	5.21	44.90
QSSVM	85.23	1.21	25.01	91.24	3.20	1656.43	91.17	0.78	138.83	92.04	0.46	2942.85
KQSSVM	80.53	21.95	0.97	92.29	5.42	4.12	91.61	0.98	4.13	**92.55**	0.59	30.52
KGSSVM	79.56	0.11	0.01	81.07	0.11	0.08	88.59	0.49	0.07	90.86	0.24	1.66
DRC-LSSVM	87.73	7.84	0.62	88.82	5.49	0.81	80.45	9.79	0.94	89.47	1.64	0.98
DRC-QSSVM	**96.51**	0.51	0.66	**97.85**	0.75	1.36	**92.54**	0.64	0.98	92.21	0.60	1.26

Table 7
Testing results on the imbalanced benchmark data sets.

Model	Careval							Heart						
	Acc(%)		AUC(%)		G-mean(%)		CPU(s)	Acc(%)		AUC(%)		G-mean(%)		CPU(s)
	mean	std	mean	std	mean	std		mean	std	mean	std	mean	std	
LSSVM	79.88	10.68	84.64	10.63	74.51	8.06	0.70	53.17	1.11	51.30	0.40	50.68	0.28	0.02
QSSVM	96.22	0.96	98.59	0.44	92.37	1.93	2.98	76.88	3.06	56.07	2.75	48.33	2.32	35.88
KQSSVM	94.72	0.55	99.11	0.28	93.45	0.91	0.70	67.09	0.60	50.11	1.01	55.10	1.53	3.10
KGSSVM	96.00	0.46	98.74	0.20	94.34	1.27	0.01	84.79	[a]0.00	57.85	[a]0.00	[a]0.00	[a]0.00	0.04
DRC-LSSVM	95.66	0.79	96.51	0.59	83.02	1.18	0.49	78.61	6.33	54.65	1.34	53.68	0.85	0.58
DRC-QSSVM	**99.65**	0.70	**99.82**	0.78	**95.97**	1.32	0.50	**85.77**	0.02	**71.32**	0.90	**66.24**	1.47	1.21

[a] A G-mean score of value 0 and std of 0 indicate the classifier simply assigns all instances to the majority class.

5. Conclusion

In this paper, we have established a novel distributionally robust chance-constrained kernel-free quadratic surface support vector machine model that can robustly conduct nonlinear classification for data sets involving stochastic uncertainties, in which only the first- and second-order moments are known a priori. SDP and SOCP reformulations of the proposed model have been derived for computational efficiency. Additionally, an explicit geometric interpretation of the conceptual distributionally robust chance constraints has been presented to show how the proposed model handle uncertain data. Our computational experiments show that the proposed model clearly outperforms the DRC-LSSVM model, the only maximum-margin SVM model explicitly using moment information for classifying uncertain data, on synthetic and public benchmark data sets.

For commonly used data sets without uncertainty involved, we design a cluster-based data-driven approach that retrieves the hidden

moment information first and enables the proposed model to leverage the moments for robust classification. This approach aids in further exploring the applicability of the proposed model. Extensive computational experiments using public benchmark data sets exhibit the surprisingly dominant performance of the proposed model over other state-of-the-art SVM models, especially for massive and/or imbalanced data sets.

Our investigation of the proposed model leads to some potential research works. First, in real-world applications, historical data may not be sufficient to capture the whole structure of the data set (Hsu, Xu, Lin, & Bell, 2022; Mi, Quan, Shi, & Wang, 2022). Collecting new data over time is necessary to adapt to changes in the data and the evolving moment information. We are interested in developing a dynamic approach to update the clustering and classification processes to extend the proposed model further. Besides, we are interested in how the proposed model performs in healthcare applications (Jiang, Han, Yu, & Ding, 2023; Naumzik, Feuerriegel, & Nielsen, 2023) with large-scale uncertain data.

Acknowledgments

This work has been sponsored by the National Science Foundation CNS-2229245, the National Natural Science Foundation of China Grants #72201052 and #72261008, Hainan Provincial Natural Science Foundation of China Grant 724RC488 and the Foundation of Yunnan Key Laboratory of Service Computing Grant #YNSC23115.

Appendix A. Proofs

A.1. Proof of Lemma 2.2

Proof. For any i, the constraints in $\mathcal{V}_i^{SDP}$ require finding β_i and $\boldsymbol{R}_i \succeq 0$ to satisfy $\frac{1}{\epsilon}\boldsymbol{\Gamma}_i \bullet \boldsymbol{R}_i - \beta_i \leqslant 0$ and the matrix inequality. Notice that finding β_i and $\boldsymbol{R}_i \succeq 0$ with $\boldsymbol{R}_i + \begin{bmatrix} \frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w} \\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^i b + \xi_i - 1 - \beta_i \end{bmatrix} \succeq 0$ is not difficult. However, it is hard to guarantee such β_i and $\boldsymbol{R}_i$ satisfies $\frac{1}{\epsilon}\boldsymbol{\Gamma}_i \bullet \boldsymbol{R}_i - \beta_i \leqslant 0$. Requiring $\mathcal{V}_i^{SDP} \neq \varnothing$ is equivalent to requiring that the optimal value of the following optimization problem is less than or equal to 0:

$$\inf \ \frac{1}{\epsilon}\boldsymbol{\Gamma}_i \bullet \boldsymbol{R}_i - \beta_i \tag{A.1a}$$

$$s.t. \ \boldsymbol{R}_i + \begin{bmatrix} \frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w} \\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^i b + \xi_i - 1 - \beta_i \end{bmatrix} \succeq 0, \tag{A.1b}$$

$$\boldsymbol{R}_i \succeq 0. \tag{A.1c}$$

Since one can easily find a proper $\beta_i \in \mathbb{R}$ and a matrix $\boldsymbol{R}_i \succ 0$ with eigenvalues large enough to satisfy $\boldsymbol{R}_i + \begin{bmatrix} \frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w} \\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^i b + \xi_i - 1 - \beta_i \end{bmatrix} \succ 0$, the Slater's condition of problem (A.1) is satisfied and the strong duality holds for (A.1) and its dual problem. Let $\bar{\boldsymbol{D}}_i = \begin{bmatrix} \boldsymbol{D}_i & \boldsymbol{d}_i \\ \boldsymbol{d}_i^{\mathrm{T}} & d_{0i} \end{bmatrix} \in \mathbb{S}_+^{n+1}$ and $\boldsymbol{C}_i \in \mathbb{S}_+^{n+1}$ be the dual variables corresponding to (A.1b) and (A.1c), respectively. Then the Lagrangian becomes

$$\begin{aligned}
&\sup_{\bar{\boldsymbol{D}}_i \geq 0, \boldsymbol{C}_i \geq 0} \inf_{\boldsymbol{R}_i, \beta_i} \mathcal{L}(\boldsymbol{R}_i, \beta_i, \bar{\boldsymbol{D}}_i, \boldsymbol{C}_i) \\
&= \sup_{\bar{\boldsymbol{D}}_i \geq 0, \boldsymbol{C}_i \geq 0} \inf_{\boldsymbol{R}_i, \beta_i} \left\{ \frac{1}{\epsilon}\boldsymbol{\Gamma}_i \bullet \boldsymbol{R}_i - \beta_i - \bar{\boldsymbol{D}}_i \bullet \left(\boldsymbol{R}_i + \begin{bmatrix} \frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w} \\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^i b + \xi_i - 1 - \beta_i \end{bmatrix} \right) \right. \\
&\qquad \left. - \boldsymbol{C}_i \bullet \boldsymbol{R}_i \right\} \\
&= \sup_{\bar{\boldsymbol{D}}_i \geq 0, \boldsymbol{C}_i \geq 0} \inf_{\boldsymbol{R}_i, \beta_i} \left\{ \boldsymbol{R}_i \bullet \left(\frac{1}{\epsilon}\boldsymbol{\Gamma}_i - \boldsymbol{C}_i - \bar{\boldsymbol{D}}_i \right) + (d_{0i} - 1)\beta_i - d_{0i}(y^i b + \xi_i - 1) \right. \\
&\qquad \left. - \frac{1}{2}y^i \boldsymbol{M} \bullet \boldsymbol{D}_i - y^i \boldsymbol{w}^{\mathrm{T}} \boldsymbol{d}_i \right\} \\
&= \sup_{\bar{\boldsymbol{D}}_i \geq 0} \begin{cases} -\frac{1}{2}y^i \boldsymbol{M} \bullet \boldsymbol{D}_i - y^i \boldsymbol{w}^{\mathrm{T}} \boldsymbol{d}_i - (y^i b + \xi_i - 1), & d_{0i} - 1 = 0, \ \frac{1}{\epsilon}\boldsymbol{\Gamma}_i - \bar{\boldsymbol{D}}_i \succeq 0, \\ -\infty, & \text{otherwise.} \end{cases}
\end{aligned}$$

Consequently, we have dual problem of (A.1):

$$\begin{aligned}
\sup \ & -\tfrac{1}{2}y^i \boldsymbol{M} \bullet \boldsymbol{D}_i - y^i \boldsymbol{w}^{\mathrm{T}} \boldsymbol{d}_i - (y^i b + \xi_i - 1) \\
s.t. \ & \begin{bmatrix} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} - \epsilon \boldsymbol{D}_i & \boldsymbol{\mu}_i - \epsilon \boldsymbol{d}_i \\ \boldsymbol{\mu}_i^{\mathrm{T}} - \epsilon \boldsymbol{d}_i^{\mathrm{T}} & 1 - \epsilon \end{bmatrix} \succeq 0, \\
& \begin{bmatrix} \boldsymbol{D}_i & \boldsymbol{d}_i \\ \boldsymbol{d}_i^{\mathrm{T}} & 1 \end{bmatrix} \succeq 0, \\
& \boldsymbol{D}_i \in \mathbb{S}^n, \ \boldsymbol{d}_i \in \mathbb{R}^n.
\end{aligned} \tag{A.2}$$

The Schur Complement Lemma implies that $\begin{bmatrix} \boldsymbol{D}_i & \boldsymbol{d}_i \\ \boldsymbol{d}_i^{\mathrm{T}} & 1 \end{bmatrix} \succeq 0 \Leftrightarrow \boldsymbol{D}_i - \boldsymbol{d}_i \boldsymbol{d}_i^{\mathrm{T}} \succeq 0$. Let $\boldsymbol{D}_i^0 = \boldsymbol{D}_i - \boldsymbol{d}_i \boldsymbol{d}_i^{\mathrm{T}}$. Substituting it into the first constraint, we have

$$\begin{aligned}
&\begin{bmatrix} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} - \epsilon \boldsymbol{D}_i & \boldsymbol{\mu}_i - \epsilon \boldsymbol{d}_i \\ \boldsymbol{\mu}_i^{\mathrm{T}} - \epsilon \boldsymbol{d}_i^{\mathrm{T}} & 1 - \epsilon \end{bmatrix} \succeq 0 \\
&\Leftrightarrow \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} - \epsilon \boldsymbol{D}_i - \frac{1}{1-\epsilon}(\boldsymbol{\mu}_i - \epsilon \boldsymbol{d}_i)(\boldsymbol{\mu}_i - \epsilon \boldsymbol{d}_i)^{\mathrm{T}} \succeq 0 \\
&\Leftrightarrow \begin{bmatrix} \boldsymbol{\Sigma}_i - \epsilon \boldsymbol{D}_i^0 & \boldsymbol{\mu}_i - \boldsymbol{d}_i \\ \boldsymbol{\mu}_i^{\mathrm{T}} - \boldsymbol{d}_i^{\mathrm{T}} & \frac{1-\epsilon}{\epsilon} \end{bmatrix} \succeq 0.
\end{aligned}$$

Therefore, we can rewrite (A.2) as

$$\begin{aligned}
\sup \ & -y^i \left(\tfrac{1}{2}\boldsymbol{d}_i^{\mathrm{T}} \boldsymbol{M} \boldsymbol{d}_i + \boldsymbol{w}^{\mathrm{T}} \boldsymbol{d}_i + b \right) - \xi_i + 1 - \tfrac{1}{2}y^i \boldsymbol{M} \bullet \boldsymbol{D}_i^0 \\
s.t. \ & \begin{bmatrix} \boldsymbol{\Sigma}_i - \epsilon \boldsymbol{D}_i^0 & \boldsymbol{\mu}_i - \boldsymbol{d}_i \\ \boldsymbol{\mu}_i^{\mathrm{T}} - \boldsymbol{d}_i^{\mathrm{T}} & \frac{1-\epsilon}{\epsilon} \end{bmatrix} \succeq 0, \\
& \boldsymbol{D}_i^0 \succeq 0, \\
& \boldsymbol{D}_i^0 \in \mathbb{S}^n, \ \boldsymbol{d}_i \in \mathbb{R}^n.
\end{aligned} \tag{A.3}$$

Remember that the strong duality holds for (A.1) and (A.3). By satisfying the requirement that the optimal value is no larger than 0, the constraints in $\mathcal{V}_i^{SDP'}$ are obtained and thus the claim follows. □

A.2. Proof of Theorem 2.4

Proof. When deriving Lemma 2.2, we notice that the SDP constraints in (7) could be obtained by requiring the optimal value of the following convex SDP less than or equal to 0:

$$\begin{aligned}
\sup \ & -\tfrac{1}{2}y^i \boldsymbol{M} \bullet \boldsymbol{D}_i - y^i \boldsymbol{w}^{\mathrm{T}} \boldsymbol{d}_i - (y^i b + \xi_i - 1) \\
s.t. \ & \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} - \epsilon \boldsymbol{D}_i - \tfrac{1}{1-\epsilon}(\boldsymbol{\mu}_i - \epsilon \boldsymbol{d}_i)(\boldsymbol{\mu}_i - \epsilon \boldsymbol{d}_i)^{\mathrm{T}} \succeq 0, \\
& \epsilon \boldsymbol{D}_i - \epsilon \boldsymbol{d}_i \boldsymbol{d}_i^{\mathrm{T}} \succeq 0, \\
& \boldsymbol{D}_i \in \mathbb{S}^n, \ \boldsymbol{d}_i \in \mathbb{R}^n.
\end{aligned} \tag{A.4}$$

Let $\boldsymbol{D}_i^* = \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} \succ 0$ and $\boldsymbol{d}_i^* = \boldsymbol{\mu}_i$, the constraints are satisfied strictly with positive definite matrices. Slater's condition is satisfied, and the strong duality holds. It is hard to give an explicit optimal solution and optimal objective value directly by solving (A.4). We consider its dual problem. Let $\boldsymbol{H}_i, \boldsymbol{G}_i \in \mathbb{S}_+^n$ be the dual variables of (A.4). The corresponding Lagrangian dual is

$$\begin{aligned}
&\inf_{\boldsymbol{H}_i \geq 0, \boldsymbol{G}_i \geq 0} \sup_{\boldsymbol{D}_i, \boldsymbol{d}_i} \mathcal{L}(\boldsymbol{D}_i, \boldsymbol{d}_i, \boldsymbol{H}_i, \boldsymbol{G}_i) \\
&= \inf_{\boldsymbol{H}_i \geq 0, \boldsymbol{G}_i \geq 0} \sup_{\boldsymbol{D}_i, \boldsymbol{d}_i} \left\{ -\tfrac{1}{2}y^i \boldsymbol{M} \bullet \boldsymbol{D}_i - y^i \boldsymbol{w}^{\mathrm{T}} \boldsymbol{d}_i - (y^i b + \xi_i - 1) \right. \\
&\qquad + \boldsymbol{H}_i \bullet \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}} - \epsilon \boldsymbol{D}_i - \tfrac{1}{1-\epsilon}(\boldsymbol{\mu}_i - \epsilon \boldsymbol{d}_i)(\boldsymbol{\mu}_i - \epsilon \boldsymbol{d}_i)^{\mathrm{T}} \right) + \boldsymbol{G}_i \bullet \left(\epsilon \boldsymbol{D}_i - \epsilon \boldsymbol{d}_i \boldsymbol{d}_i^{\mathrm{T}} \right) \Big\} \\
&= \inf_{\boldsymbol{H}_i \geq 0, \boldsymbol{G}_i \geq 0} \sup_{\boldsymbol{D}_i, \boldsymbol{d}_i} \left\{ -(y^i b + \xi_i - 1) + \boldsymbol{H}_i \bullet \boldsymbol{\Sigma}_i - \tfrac{\epsilon}{1-\epsilon}\boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{H}_i \boldsymbol{\mu}_i \right. \\
&\qquad + \boldsymbol{D}_i \bullet \left(-\tfrac{1}{2}y^i \boldsymbol{M} - \epsilon \boldsymbol{H}_i + \epsilon \boldsymbol{G}_i \right) - \boldsymbol{d}_i^{\mathrm{T}} \left(\tfrac{\epsilon^2}{1-\epsilon}\boldsymbol{H}_i + \epsilon \boldsymbol{G}_i \right) \boldsymbol{d}_i + \left(-y^i \boldsymbol{w} + \tfrac{2\epsilon}{1-\epsilon}\boldsymbol{H}_i \boldsymbol{\mu}_i \right)^{\mathrm{T}} \boldsymbol{d}_i \Big\}.
\end{aligned}$$

The dual function is finite if and only if $-\frac{1}{2}y^i \boldsymbol{M} - \epsilon \boldsymbol{H}_i - \epsilon \boldsymbol{G}_i = 0$, which gives $\epsilon \boldsymbol{G}_i = \frac{1}{2}y^i \boldsymbol{M} + \epsilon \boldsymbol{H}_i \succeq 0$. Substitute this into the above, we get

$$\begin{aligned}
&\inf_{\boldsymbol{H}_i \geq 0, \boldsymbol{G}_i \geq 0} \sup_{\boldsymbol{D}_i, \boldsymbol{d}_i} \mathcal{L}(\boldsymbol{D}_i, \boldsymbol{d}_i, \boldsymbol{H}_i, \boldsymbol{G}_i) \\
&= \inf_{\boldsymbol{H}_i \geq 0} \sup_{\boldsymbol{d}_i} \begin{cases} -(y^i b + \xi_i - 1) + \boldsymbol{H}_i \bullet \boldsymbol{\Sigma}_i - \frac{\epsilon}{1-\epsilon}\boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{H}_i \boldsymbol{\mu}_i + q(\boldsymbol{d}_i), & \text{if } \frac{1}{2}y^i \boldsymbol{M} + \epsilon \boldsymbol{H}_i \succeq 0, \\ +\infty, & \text{otherwise,} \end{cases}
\end{aligned}$$

where $q(\boldsymbol{d}_i) = -\boldsymbol{d}_i^{\mathrm{T}}(\frac{\epsilon}{1-\epsilon}\boldsymbol{H}_i + \frac{1}{2}y^i\boldsymbol{M})\boldsymbol{d}_i + (-y^i\boldsymbol{w} + \frac{2\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{d}_i$. Since the dual variable $\boldsymbol{H}_i \succeq 0$, we have $\frac{\epsilon}{1-\epsilon}\boldsymbol{H}_i + \frac{1}{2}y^i\boldsymbol{M} \succeq \epsilon\boldsymbol{H}_i + \frac{1}{2}y^i\boldsymbol{M} \succeq 0$ by $0 < \epsilon < 1$. Thus, $q(\boldsymbol{d}_i)$ is a concave function of $\boldsymbol{d}_i$ since $\frac{\epsilon}{1-\epsilon}\boldsymbol{H}_i + \frac{1}{2}y^i\boldsymbol{M} \succeq 0$. To make it strictly concave, we can add $\eta\boldsymbol{I}_n$ to $\frac{\epsilon}{1-\epsilon}\boldsymbol{H}_i + \frac{1}{2}y^i\boldsymbol{M}$ such that $\boldsymbol{A}_i(\eta) \triangleq \frac{\epsilon}{1-\epsilon}\boldsymbol{H}_i + \frac{1}{2}y^i\boldsymbol{M} + \eta\boldsymbol{I}_n \succ 0$ and $\lim_{\eta\to 0}\boldsymbol{A}_i(\eta) = \frac{\epsilon}{1-\epsilon}\boldsymbol{H}_i + \frac{1}{2}y^i\boldsymbol{M}$. Then solving $\nabla_{\boldsymbol{d}_i} q(\boldsymbol{d}) = -2\boldsymbol{A}_i(\eta)\boldsymbol{d}_i - (y^i\boldsymbol{w} - \frac{2\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i) = 0$, we have $\boldsymbol{d} = -\frac{1}{2}\boldsymbol{A}_i(\eta)^{-1}(y^i\boldsymbol{w} - \frac{2\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i)$. Then we get the dual function as follows:

$$\begin{aligned} g(\boldsymbol{H}_i) &= -(y^i b + \xi_i - 1) + \boldsymbol{H}_i \bullet \boldsymbol{\Sigma}_i - \tfrac{\epsilon}{1-\epsilon}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{\mu}_i \\ &\quad + \tfrac{1}{4}(y^i\boldsymbol{w} - \tfrac{2\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(y^i\boldsymbol{w} - \tfrac{2\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i) \\ &= -(y^i b + \xi_i - 1) + \boldsymbol{H}_i \bullet \boldsymbol{\Sigma}_i - \tfrac{\epsilon}{1-\epsilon}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{\mu}_i + \tfrac{1}{4}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}\boldsymbol{w} \\ &\quad - y^i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\tfrac{\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i) + (\tfrac{\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\tfrac{\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i). \end{aligned}$$

For the last two terms, We have $y^i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\frac{\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i) = y^i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_i - y^i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\frac{1}{2}y^i\boldsymbol{M}\boldsymbol{\mu}_i + \eta\boldsymbol{\mu}_i)$, and $(\frac{\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\frac{\epsilon}{1-\epsilon}\boldsymbol{H}_i\boldsymbol{\mu}_i) = \frac{\epsilon}{1-\epsilon}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{\mu}_i - \frac{1}{2}y^i\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i - \eta\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_i + (\frac{1}{2}y^i\boldsymbol{M}\boldsymbol{\mu}_i + \eta\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\frac{1}{2}y^i\boldsymbol{M}\boldsymbol{\mu}_i + \eta\boldsymbol{\mu}_i)$. Hence, we can derive

$$\begin{aligned} g(\boldsymbol{H}_i) &= \tfrac{1}{4}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}\boldsymbol{w} + (\tfrac{1}{2}y^i\boldsymbol{M}\boldsymbol{\mu}_i + \eta\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\tfrac{1}{2}y^i\boldsymbol{M}\boldsymbol{\mu}_i + \eta\boldsymbol{\mu}_i) \\ &\quad + y^i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\tfrac{1}{2}y^i\boldsymbol{M}\boldsymbol{\mu}_i + \eta\boldsymbol{\mu}_i) \\ &\quad + \boldsymbol{H}_i \bullet \boldsymbol{\Sigma}_i - \tfrac{1}{2}y^i\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i - y^i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_i - \eta\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_i - (y^i b + \xi_i - 1) \\ &= (\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i) \\ &\quad + \boldsymbol{H}_i \bullet \boldsymbol{\Sigma}_i - \tfrac{1}{2}y^i\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i - y^i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_i - \eta\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_i - (y^i b + \xi_i - 1). \end{aligned}$$

The dual problem of (A.4) is followed by

$$\begin{aligned} \inf \quad & g(\boldsymbol{H}_i) = (\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i) \\ &\qquad + \boldsymbol{H}_i \bullet \boldsymbol{\Sigma}_i - \tfrac{1}{2}y^i\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i - y^i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_i - \eta\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_i - (y^i b + \xi_i - 1) \\ s.t. \quad & \epsilon\boldsymbol{H}_i + \tfrac{1}{2}y^i\boldsymbol{M} \succeq 0, \\ & \boldsymbol{H}_i \succeq 0. \end{aligned} \tag{A.5}$$

A bounded unconstrained convex program (A.5) is obtained. And $\nabla g(\boldsymbol{H}_i) = 0$ implies that

$$\boldsymbol{A}_i(\eta)^{-1}(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1} = \frac{1-\epsilon}{\epsilon}\boldsymbol{\Sigma}_i. \tag{A.6}$$

Multiply $(\frac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)^{\mathrm{T}}$ on the left hand side and $(\frac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)$ on the right hand side of (A.6), we have

$$\begin{aligned} & \left((\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)\right)^2 \\ = \quad & \tfrac{1-\epsilon}{\epsilon}(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i) \\ \Rightarrow \quad & (\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i) \\ = \quad & \sqrt{\tfrac{1-\epsilon}{\epsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)\|_2. \end{aligned} \tag{A.7}$$

Multiply $\boldsymbol{A}_i(\eta)$ on the left hand side of (A.6) and take the trace, we have

$$\begin{aligned} & (\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)^{\mathrm{T}} \bullet \boldsymbol{A}_i(\eta)^{-1} \\ & = (\frac{\epsilon}{1-\epsilon}\boldsymbol{H}_i + \frac{1}{2}y^i\boldsymbol{M} + \eta\boldsymbol{I}_n) \bullet \frac{1-\epsilon}{\epsilon}\boldsymbol{\Sigma}_i \\ \Rightarrow & (\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{A}_i(\eta)^{-1}(\tfrac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i) \\ & = \boldsymbol{H}_i \bullet \boldsymbol{\Sigma}_i + \frac{1-\epsilon}{2\epsilon}y^i\boldsymbol{M} \bullet \boldsymbol{\Sigma}_i + \eta\frac{1-\epsilon}{\epsilon}Trace(\boldsymbol{\Sigma}_i). \end{aligned} \tag{A.8}$$

By (A.7) and (A.8), we have $\inf g(\boldsymbol{H}_i) = -\frac{1}{2}y^i\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i - y^i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_i + \sqrt{\frac{1-\epsilon}{\epsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\frac{1}{2}y^i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}) + \eta\boldsymbol{\mu}_i)\|_2 - \frac{1-\epsilon}{2\epsilon}y^i\boldsymbol{M} \bullet \boldsymbol{\Sigma}_i - \eta\frac{1-\epsilon}{\epsilon}Trace(\boldsymbol{\Sigma}_i) - \eta\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_i - (y^i b + \xi_i - 1)$, and $\lim_{\eta\to 0}\inf g(\boldsymbol{H}_i) = -\frac{1}{2}y^i\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i - y^i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_i + \sqrt{\frac{1-\epsilon}{\epsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})\|_2 - \frac{1-\epsilon}{2\epsilon}y^i\boldsymbol{M} \bullet \boldsymbol{\Sigma}_i - (y^i b + \xi_i - 1)$. Since the strong duality holds, we require that $\lim_{\eta\to 0}\inf g(\boldsymbol{H}_i) \leqslant 0$ which yields SOC constraints $y^i(\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{w} + b) \geqslant 1 - \xi_i + \sqrt{\frac{1-\epsilon}{\epsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})\|_2 - \frac{1-\epsilon}{2\epsilon}y^i\boldsymbol{M} \bullet \boldsymbol{\Sigma}_i$. This completes the proof. □

A.3. Proof of Lemma 2.5

Proof. In $\mathcal{V}_i^E$, for any $\boldsymbol{x}^i \in \mathcal{E}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \frac{1-\epsilon}{\epsilon})$, we require that $y^i\left(\frac{1}{2}(\boldsymbol{x}^i)^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}^i + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^i + b\right) \geqslant 1 - \xi_i$ which is equivalent to

$$\begin{aligned} y^i\left(\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{w} + b\right) \geqslant 1 - \xi_i - \sqrt{\frac{1-\epsilon}{\epsilon}}y^i\left(\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})\right)^{\mathrm{T}}\boldsymbol{v} \\ - \frac{1-\epsilon}{2\epsilon}y^i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{v}, \end{aligned} \tag{A.9}$$

for any $\boldsymbol{v} \in \mathbb{R}^n$ with $\|\boldsymbol{v}\|_2 \leqslant 1$. To eliminate $\boldsymbol{v}$, we need to know the maximum of the RHS of the above inequality. We need to solve the following subproblem:

$$\begin{aligned} \inf_{\boldsymbol{v}\in\mathbb{R}^n} \quad & \sqrt{\tfrac{1-\epsilon}{\epsilon}}y^i(\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}))^{\mathrm{T}}\boldsymbol{v} + \tfrac{1-\epsilon}{2\epsilon}y^i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{v} \\ s.t. \quad & \|\boldsymbol{v}\|_2 \leqslant 1. \end{aligned} \tag{A.10}$$

Problem (A.10) is a typical trust-region subproblem. The dual problem of (A.10) can be derived as follows:

$$\begin{aligned} \sup_{\lambda\in\mathbb{R}^{++}} \quad & \phi(\lambda) \\ s.t. \quad & \tfrac{1-\epsilon}{\epsilon}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}_i^{\frac{1}{2}} + \lambda\boldsymbol{I} \succ 0, \end{aligned} \tag{A.11}$$

where $\phi(\lambda) = -\frac{1}{2}\sqrt{\frac{1-\epsilon}{\epsilon}}(\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w}))^{\mathrm{T}}\sqrt{\frac{1-\epsilon}{\epsilon}}(\frac{1-\epsilon}{\epsilon}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}_i^{\frac{1}{2}} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})) + \lambda$. A pair $(\boldsymbol{v}^*, \lambda^*)$ provides primal and dual optimal solutions if and only if

$$\begin{cases} \left(\frac{1-\epsilon}{\epsilon}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}_i^{\frac{1}{2}} + \lambda\boldsymbol{I}\right)\boldsymbol{v}^* = -\sqrt{\frac{1-\epsilon}{\epsilon}}\left(\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})\right) \\ \lambda^*(\|\boldsymbol{v}^*\|_2 - 1) = 0, \|\boldsymbol{v}^*\|_2 \leqslant 1 \\ \frac{1-\epsilon}{\epsilon}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}_i^{\frac{1}{2}} + \lambda^*\boldsymbol{I} \succ 0, \lambda^* > 0. \end{cases} \tag{A.12}$$

Let $\hat{\boldsymbol{v}} = -y^i\frac{\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})}{\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})\|_2}$. We have $\|\hat{\boldsymbol{v}}\|_2 = 1$, thus $\hat{\boldsymbol{v}}$ is a feasible solution of the primal problem (A.10). The objective value respect to $\hat{\boldsymbol{v}}$ is $-\sqrt{\frac{1-\epsilon}{\epsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})\|_2 + y^i\frac{1-\epsilon}{2\epsilon}\hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\hat{\boldsymbol{v}}$. For the second term of the objective value, we have

$$\begin{aligned} \hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\hat{\boldsymbol{v}} &= \frac{(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{M}\boldsymbol{\Sigma}_i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})}{(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}_i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})} \\ &= \frac{\boldsymbol{M}\boldsymbol{\Sigma}_i \bullet (\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}_i}{(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}_i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})}. \end{aligned}$$

According to Von Neumann's trace inequality, we have that $\theta_{\min}(\boldsymbol{M}\boldsymbol{\Sigma}_i)Trace((\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}_i) \leqslant \boldsymbol{M}\boldsymbol{\Sigma}_i \bullet (\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}_i \leqslant \theta_{\max}(\boldsymbol{M}\boldsymbol{\Sigma}_i)Trace((\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}_i)$, where θ denotes the eigenvalue of $\boldsymbol{M}\boldsymbol{\Sigma}_i$. Since $Trace((\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}_i) = (\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})^{\mathrm{T}}\boldsymbol{\Sigma}_i(\boldsymbol{M}\boldsymbol{\mu}_i + \boldsymbol{w})$, then $\theta_{\min}(\boldsymbol{M}\boldsymbol{\Sigma}_i) \leqslant \hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\hat{\boldsymbol{v}} \leqslant \theta_{\max}(\boldsymbol{M}\boldsymbol{\Sigma}_i)$. If $\boldsymbol{M}$ is positive semi-definite, we have $\hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\hat{\boldsymbol{v}} \leqslant Trace(\boldsymbol{M}\boldsymbol{\Sigma}_i)$. In this case, substituting the corresponding objective value make the inequality (A.9) become the constraint of $\mathcal{V}_i^{SOC}$ in (15) for each i. However, we can verify that $\hat{\boldsymbol{v}}$ does not satisfy the optimal conditions (A.12), which means that $\mathcal{V}_i^E$ is a conservative cut of $\mathcal{V}_i^{SOC}$. Hence, we have $\mathcal{V}_i^E \subset \mathcal{V}_i^{SOC}$.

For the linear case when $\boldsymbol{M} = 0$, the problem (A.10) becomes $\inf_{\|\boldsymbol{v}\|_2\leqslant 1}\sqrt{\frac{1-\epsilon}{\epsilon}}y^i(\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{w})^{\mathrm{T}}\boldsymbol{v}$. The optimal solution is that $\boldsymbol{v}^* = -y^i\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{w}/\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{w}\|_2$. Consequently, the constraint (A.9) becomes $y^i\left(\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{w} + b\right) \geqslant 1 - \xi_i + \sqrt{\frac{1-\epsilon}{\epsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{w}\|_2$, which is equivalent to $y^i\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} + b\right) \geqslant 1 - \xi_i$, $\forall \boldsymbol{x} \in \boldsymbol{V}_i^E$ by utilizing the multivariate Chebyshev inequality (Bertsimas &

Table B.8
Testing results on "elliptic" synthetic data sets by DRC-QSSVM.

Data		Syn-Ellip-4d-50				Syn-Ellip-8d-50				Syn-Ellip-16d-50			
ϵ		0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
Acc(%)		99.52	99.60	99.20	98.64	94.04	99.88	99.52	99.08	98.88	99.84	99.92	99.68
PrAcc(%)		97.17	97.97	96.46	95.32	87.12	88.74	88.53	88.31	78.67	84.76	86.94	87.11
CPU time (s)	SOCP	4.66	4.69	4.71	3.78	4.78	4.76	4.72	3.85	5.72	5.71	5.98	3.85
	SDP	6.87	7.14	7.13	5.71	9.97	9.59	9.65	8.28	34.96	30.90	29.54	21.77
Data		Syn-Ellip-4d-100				Syn-Ellip-8d-100				Syn-Ellip-16d-100			
ϵ		0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
Acc(%)		99.40	99.64	99.56	99.60	100.00	100.00	100.00	98.80	98.60	99.64	99.64	99.64
PrAcc(%)		97.99	98.43	98.21	98.27	92.38	92.46	92.47	89.49	84.49	89.22	89.83	90.72
CPU time (s)	SOCP	7.19	7.37	7.38	5.46	5.98	5.90	6.14	3.67	13.49	13.90	13.92	7.39
	SDP	11.78	12.99	12.93	8.68	18.04	16.10	16.16	11.46	34.96	30.90	29.54	21.77

Popescu, 2005): for an arbitrary closed convex set S, $\mathbb{P}(\boldsymbol{x} \in S) \leqslant \frac{1}{1+c^2}$, where $c^2 = \inf_{\boldsymbol{x}\in S} (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. Notice that the set $\{\boldsymbol{x}|y^i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\right) \geqslant 1-\xi_i\}$ is a convex set for each i. Hence, we have $\mathcal{V}_i^{SOC} = \mathcal{V}_i^{E}$. This completes the proof. □

Appendix B. Auxiliary information

B.1. Moments uncertainty

The uncertainty of moments mentioned in Remark 3.1 adopts a general bounded set (Delage & Ye, 2010):

$$\left(\mathbb{E}_{F_i}[\tilde{\boldsymbol{x}}^i]-\boldsymbol{\mu}_i\right)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\left(\mathbb{E}_{F_i}[\tilde{\boldsymbol{x}}^i]-\boldsymbol{\mu}_i\right) \leqslant \omega_1, \tag{B.1a}$$

$$\mathbb{E}_{F_i}\left[(\tilde{\boldsymbol{x}}^i-\boldsymbol{\mu}_i)(\tilde{\boldsymbol{x}}^i-\boldsymbol{\mu}_i)^{\mathrm{T}}\right] \preceq \omega_2\boldsymbol{\Sigma}_i, \tag{B.1b}$$

where the parameters $\omega_1 \geqslant 0$ and $\omega_2 \geqslant 1$ provide natural means of quantifying one's confidence in $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. Constraint (B.1a) provides an ellipsoidal uncertainty of $\boldsymbol{\mu}_i$, and constraint (B.1b) assumes that $\boldsymbol{\Sigma}_i$ lies in a positive semidefinite cone. In what follows, to overcome the possible estimation errors of moments, we consider the ambiguity set for each i,

$$D_i^{DY}(\tilde{\boldsymbol{x}}^i;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\omega_1,\omega_2) \triangleq \left\{F_i \in \mathcal{M}(\Xi_i,\mathcal{F}_i) \;\middle|\; \begin{array}{l} \mathbb{P}(\tilde{\boldsymbol{x}}^i \in \Xi_i)=1, \\ \left(\mathbb{E}_{F_i}[\tilde{\boldsymbol{x}}^i]-\boldsymbol{\mu}_i\right)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\left(\mathbb{E}_{F_i}[\tilde{\boldsymbol{x}}^i]-\boldsymbol{\mu}_i\right) \leqslant \omega_1, \\ \mathbb{E}_{F_i}\left[(\tilde{\boldsymbol{x}}^i-\boldsymbol{\mu}_i)(\tilde{\boldsymbol{x}}^i-\boldsymbol{\mu}_i)^{\mathrm{T}}\right] \preceq \omega_2\boldsymbol{\Sigma}_i \end{array}\right\}. \tag{B.2}$$

Let $D^{DY} \triangleq \bigcup_i D_i^{DY}$. Similarly, we can circumvent the difficulty of solving distributionally robust chance-constrained problems by duality theory. An SDP model can be obtained accordingly.

Lemma B.1. *Suppose that $\omega_1 \geqslant 0$, $\omega_2 \geqslant 1$ and $\boldsymbol{\Sigma}_i \succ 0$ for any i. Then, the DRC-QSSVM model under the ambiguity set D^{DY} can be equivalently reformulated as the following SDP model:*

$$\begin{array}{lll} \min & \sum_{i=1}^N \eta_i & \\ s.t. & \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{M}\boldsymbol{\mu}_i+\boldsymbol{w} \\ (\boldsymbol{M}\boldsymbol{\mu}_i+\boldsymbol{w})^{\mathrm{T}} & -\frac{C}{N}\xi_i+\eta_i \end{bmatrix} \succeq 0, & i=1,\ldots,N, \\ & \beta_i - \frac{1}{\epsilon}\boldsymbol{\Gamma}_i^0 \bullet \boldsymbol{R}_i \geqslant \frac{1}{\epsilon}\sqrt{\omega_1}\|\boldsymbol{\Sigma}_i^0\boldsymbol{R}_i\boldsymbol{\mu}_i^0\|, & i=1,\ldots,N, \\ & \boldsymbol{R}_i + \begin{bmatrix} \frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w} \\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^ib+\xi_i-1-\beta_i \end{bmatrix} \succeq 0, & i=1,\ldots,N, \\ & \boldsymbol{R}_i \succeq 0, & i=1,\ldots,N, \\ & \boldsymbol{M}\in\mathbb{S}^n,\ \boldsymbol{w}\in\mathbb{R}^n,\ b\in\mathbb{R},\ \boldsymbol{\xi}\in\mathbb{R}_+^N,\ \boldsymbol{\beta},\ \boldsymbol{\eta}\in\mathbb{R}^N,\ \boldsymbol{R}_i\in\mathbb{S}^{n+1}, & i=1,\ldots,N, \end{array} \tag{B.3}$$

where $\boldsymbol{\Gamma}_i^0 = \begin{bmatrix} \omega_2\boldsymbol{\Sigma}_i+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^{\mathrm{T}} & 1 \end{bmatrix}$, $\boldsymbol{\Sigma}_i^0 = [\boldsymbol{\Sigma}_i^{\frac{1}{2}}\ \mathbf{1}_n] \in \mathbb{R}^{n\times(n+1)}$, *and* $\boldsymbol{\mu}_i^0 = [\boldsymbol{\mu}_i\ 0]^{\mathrm{T}} \in \mathbb{R}^{n+1}$. *The above result holds for the linear case when* $\boldsymbol{M}=0$.

Proof. Let $\varphi_i^{DY} \triangleq \sup_{F\in D_i^{DY}} \mathbb{P}\left\{y^i\left(\frac{1}{2}(\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{M}\tilde{\boldsymbol{x}}^i+\boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}^i+b\right) \leqslant 1-\xi_i\right\} = \sup_{F\in D_i^{DY}} \mathbb{E}_F[\mathbb{1}(\tilde{\boldsymbol{x}}^i)]$, for $i=1,\ldots,N$. From Lemma 1 in Delage and Ye (2010), φ_i^{DY} must be equal to the optimal value of the problem:

$$\inf\ r_i+t_i \tag{B.4a}$$

$$s.t.\ r_i \geqslant \mathbb{1}(\tilde{\boldsymbol{x}}^i) - (\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{Q}_i\tilde{\boldsymbol{x}}^i - (\tilde{\boldsymbol{x}}^i)^{\mathrm{T}}\boldsymbol{q}_i \quad \forall\ \tilde{\boldsymbol{x}}^i\in\xi_i, \tag{B.4b}$$

$$t_i \geqslant (\omega_2\boldsymbol{\Sigma}_i+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}})\bullet\boldsymbol{Q}_i + \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{q}_i + \sqrt{\omega_1}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}(\boldsymbol{q}_i+2\boldsymbol{Q}_i\boldsymbol{\mu}_i)\|, \tag{B.4c}$$

$$\boldsymbol{Q}_i \succeq 0, \tag{B.4d}$$

$$r_i,\ t_i\in\mathbb{R},\ \boldsymbol{Q}_i\in\mathbb{S}^n,\ \boldsymbol{q}_i\in\mathbb{R}^n. \tag{B.4e}$$

Let $\boldsymbol{N}_i = \begin{bmatrix} \boldsymbol{Q}_i & \frac{1}{2}\boldsymbol{q}_i \\ \frac{1}{2}\boldsymbol{q}_i^{\mathrm{T}} & r_i \end{bmatrix}$ and $\boldsymbol{\Gamma}_i^0 = \begin{bmatrix} \omega_2\boldsymbol{\Sigma}_i+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^{\mathrm{T}} & 1 \end{bmatrix}$. Using the same approach in the proof of Theorem 2.1, the constraint (B.4b) is equivalent to $\boldsymbol{N}_i + \alpha_i\begin{bmatrix} \frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w} \\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^ib+\xi_i-1-\frac{1}{\alpha_i} \end{bmatrix} \succeq 0,\ \alpha_i>0$. Denote $\boldsymbol{\Sigma}_i^0 = [\boldsymbol{\Sigma}_i^{\frac{1}{2}}\ \mathbf{1}_n] \in \mathbb{R}^{n\times(n+1)}$, and $\boldsymbol{\mu}_i^0 = [\boldsymbol{\mu}_i\ 0]^{\mathrm{T}} \in \mathbb{R}^{n+1}$. Reformulate (B.4a) as $r_i+t_i \geqslant \boldsymbol{\Gamma}_i^0\bullet\boldsymbol{N}_i + \sqrt{\omega_1}\|\boldsymbol{\Sigma}_i^0\boldsymbol{N}_i\boldsymbol{\mu}_i^0\|$. Hence, we can rewrite (B.4) as follows:

$$\begin{array}{ll} \inf & \boldsymbol{\Gamma}_i^0\bullet\boldsymbol{N}_i + \sqrt{\omega_1}\|\boldsymbol{\Sigma}_i^0\boldsymbol{N}_i\boldsymbol{\mu}_i^0\| \\ s.t. & \boldsymbol{N}_i + \alpha_i\begin{bmatrix} \frac{1}{2}y^i\boldsymbol{M} & \frac{1}{2}y^i\boldsymbol{w} \\ \frac{1}{2}y^i\boldsymbol{w}^{\mathrm{T}} & y^ib+\xi_i-1-\frac{1}{\alpha_i} \end{bmatrix} \succeq 0, \\ & \boldsymbol{N}_i \succeq 0, \\ & \alpha_i\in\mathbb{R}^{++},\ \boldsymbol{N}_i\in\mathbb{S}^{n+1}. \end{array}$$

Since $\varphi_i^{DY} \leqslant \epsilon$, by strong duality, we have $\boldsymbol{\Gamma}_i^0\bullet\boldsymbol{N}_i + \sqrt{\omega_1}\|\boldsymbol{\Sigma}_i^0\boldsymbol{N}_i\boldsymbol{\mu}_i^0\| \leqslant \epsilon$. The rest of the proof is similar to the proof of Theorem 2.1 and can be easily followed to complete the claim. □

Notice that $D_i^{DY}(\tilde{\boldsymbol{x}}^i;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,0,1)$ relates closely to $D_i(\tilde{\boldsymbol{x}}^i;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, and in this event, we find that the model (B.3) reduces to the model (12).

B.2. Experimental results

See Tables B.8 and B.9.

Table B.9
Testing results on "parabolic" synthetic data sets by DRC-QSSVM.

Data		Syn-Para-4d-50				Syn-Para-8d-50				Syn-Para-16d-50			
ϵ		0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
Acc(%)		97.36	98.88	98.68	98.12	98.80	98.52	98.16	98.12	97.90	98.42	96.70	98.18
PrAcc(%)		93.85	98.21	98.01	97.31	97.25	97.24	96.97	96.96	86.90	90.03	88.03	88.30
CPU time (s)	SOCP	4.06	3.79	3.87	3.69	3.83	4.01	3.87	3.93	5.72	5.73	5.75	5.46
	SDP	6.28	6.06	5.96	6.57	8.76	8.22	8.32	8.49	33.49	31.97	30.44	30.31
Data		Syn-Para-4d-100				Syn-Para-8d-100				Syn-Para-16d-100			
ϵ		0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
Acc(%)		99.40	99.32	99.20	99.04	98.88	98.76	98.48	98.32	98.44	97.00	96.06	96.66
PrAcc(%)		99.06	98.97	98.75	98.54	97.85	97.69	97.39	97.22	95.20	91.78	92.59	92.95
CPU time (s)	SOCP	6.97	6.89	6.80	6.60	5.62	23.74	22.63	5.58	10.21	10.42	10.27	10.26
	SDP	9.98	9.99	9.68	9.66	16.93	15.08	15.38	15.88	153.89	140.78	132.56	140.67

References

Ben-Tal, A., Bhadra, S., Bhattacharyya, C., & Nath, J. S. (2011). Chance constrained uncertain classification via robust optimization. *Mathematical Programming*, *127*(1), 145–173.

Bertsimas, D., Dunn, J., Pawlowski, C., & Zhuo, Y. D. (2019). Robust classification. *INFORMS Journal on Optimization*, *1*(1), 2–34.

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

Bhattacharyya, C., Grate, L., Jordan, M. I., Ghaoui, L. E., & Mian, I. S. (2004). Robust sparse hyperplane classifiers: Application to uncertain molecular profiling data. *Journal of Computational Biology*, *11*(6), 1073–1089.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27.

Chen, R., & Paschalidis, I. C. (2020). Distributionally robust learning. *Foundations and Trends® in Optimization*, *4*(1–2), 1–243.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Dagher, I. (2008). Quadratic kernel-free non-linear support vector machine. *Journal of Global Optimization*, *41*(1), 15–30.

Delage, E., & Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, *58*(3), 595–612.

Gao, Z., Fang, S.-C., Luo, J., & Medhin, N. (2021). A kernel-free double well potential support vector machine with applications. *European Journal of Operational Research*, *290*(1), 248–262.

Goldfarb, D., & Iyengar, G. (2003). Robust convex quadratically constrained programs. *Mathematical Programming*, *97*(3), 495–515.

Hsu, W.-K., Xu, J., Lin, X., & Bell, M. R. (2022). Integrated online learning and adaptive control in queueing systems with uncertain payoffs. *Operations Research*, *70*(2), 1166–1181.

Huang, G., Song, S., Gupta, J. N., & Wu, C. (2013). A second order cone programming approach for semi-supervised learning. *Pattern Recognition*, *46*(12), 3548–3558.

Jiang, R., Han, S., Yu, Y., & Ding, W. (2023). An access control model for medical big data based on clustering and risk. *Information Sciences*, *621*, 691–707.

Jiménez-Cordero, A., Morales, J. M., & Pineda, S. (2021). A novel embedded min-max approach for feature selection in nonlinear support vector machine classification. *European Journal of Operational Research*, *293*(1), 24–35.

Khanjani-Shiraz, R., Babapour-Azar, A., Hosseini-Nodeh, Z., & Pardalos, P. M. (2023). Distributionally robust joint chance-constrained support vector machines. *Optimization Letters*, *17*(2), 299–332.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., & Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. In N. Serguei, S. Douglas, & H. J. Greenberg (Eds.), *Operations research & management science in the age of analytics* (pp. 130–166). INFORMS.

Lin, F., Fang, X., & Gao, Z. (2022). Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control & Optimization*, *12*(1), 159–212.

Luo, J., Fang, S.-C., Deng, Z., & Guo, X. (2016). Soft quadratic surface support vector machine for binary classification. *Asia-Pacific Journal of Operational Research*, *33*(6), Article 1650046.

Luo, J., Fang, S.-C., Deng, Z., & Tian, Y. (2022). Robust kernel-free support vector regression based on optimal margin distribution. *Knowledge-Based Systems*, *253*, Article 109477.

Luo, J., Yan, X., & Tian, Y. (2020). Unsupervised quadratic surface support vector machine with application to credit risk assessment. *European Journal of Operational Research*, *280*(3), 1008–1017.

Ma, Q., & Wang, Y. (2021). Distributionally robust chance constrained svm model with l_2-wasserstein distance. *Journal of Industrial and Management Optimization*.

Mi, Y., Quan, P., Shi, Y., & Wang, Z. (2022). Concept-cognitive computing system for dynamic classification. *European Journal of Operational Research*, *301*(1), 287–299.

Naumzik, C., Feuerriegel, S., & Nielsen, A. M. (2023). Data-driven dynamic treatment planning for chronic diseases. *European Journal of Operational Research*, *305*(2), 853–867.

Peng, S., Gianpiero, C., & Zhihua, A.-Z. (2023). Chance constrained conic-segmentation support vector machine with uncertain data. *Annals of Mathematics and Artificial Intelligence*, 1–23.

Shivaswamy, P. K., Bhattacharyya, C., & Smola, A. J. (2006). Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, *7*(47), 1283–1314.

Singla, M., Ghosh, D., & Shukla, K. (2020). A survey of robust optimization based machine learning with special reference to support vector machines. *International Journal of Machine Learning and Cybernetics*, *11*(7), 1359–1385.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(04), 687–719.

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*, 429–441.

Toh, K.-C., Todd, M. J., & Tütüncü, R. H. (1999). SDPT3 – a Matlab software package for semidefinite programming, Version 1.3. *Optimization Methods & Software*, *11*(1–4), 545–581.

Trafalis, T. B., & Gilbert, R. C. (2006). Robust classification and regression using support vector machines. *European Journal of Operational Research*, *173*(3), 893–909.

Vassilvitskii, S., & Arthur, D. (2006). K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1027–1035).

Wang, X., Fan, N., & Pardalos, P. M. (2017). Stochastic subgradient descent method for large-scale robust chance-constrained support vector machines. *Optimization Letters*, *11*, 1013–1024.

Wang, X., Fan, N., & Pardalos, P. M. (2018). Robust chance-constrained support vector machines with second-order moment information. *Annals of Operations Research*, *263*(1), 45–68.

Wang, X., & Pardalos, P. M. (2014). A survey of support vector machines with uncertainties. *Annals of Data Science*, *1*(3–4), 293–309.

Zhou, Z. (2021). *Machine learning*. Springer Nature (Chapter 6).