# Robust and Distributionally Robust Optimization Models for Linear Support Vector Machine

Daniel Faccini[a], Francesca Maggioni[a,*], Florian A. Potra[b]

[a] *Department of Management, Information and Production Engineering, University of Bergamo*
*Viale G. Marconi 5, 24044, Dalmine, Italy*

[b] *Department of Mathematics and Statistics, University of Maryland*
*Baltimore County, USA*

**Abstract**

In this paper we present novel data-driven optimization models for Support Vector Machines (SVM), with the aim of linearly separating two sets of points that have non-disjoint convex closures. Traditional classification algorithms assume that the training data points are always known exactly. However, real-life data are often subject to noise. To handle such uncertainty, we formulate robust models with uncertainty sets in the form of hyperrectangles or hyperellipsoids, and propose a moment-based distributionally robust optimization model enforcing limits on first-order deviations along principal directions. All the formulations reduce to convex programs. The efficiency of the new classifiers is evaluated on real-world databases. Experiments show that robust classifiers are especially beneficial for data sets with a small number of observations. As the dimension of the data sets increases, features behavior is gradually learned and higher levels of out-of-sample accuracy can be achieved via the considered distributionally robust optimization method. The proposed formulations, overall, allow finding a trade-off between increasing the average performance accuracy and protecting against uncertainty, with respect to deterministic approaches.

*Keywords:* Machine Learning; Support Vector Machine; Robust Optimization; Distributionally Robust Optimization.

## 1. Introduction

Binary pattern separation is one of the main *Machine Learning* (ML) tasks [51]. Its aim is to classify observations into one of two classes and it is a critical problem in many practical application fields, such as robotics [22], environmental engineering [26, 73, 74], nutrition [87],

[*] Corresponding author

*Email addresses:* Daniel.Faccini@unibg.it (Daniel Faccini), Francesca.Maggioni@unibg.it (Francesca Maggioni), Potra@math.umbc.edu (Florian A. Potra)

neural and medical image analysis [118] and computer security [17]. From the ML standpoint, a great variety of algorithms have been devised to address the classification problem: *Decision Trees* (DT) [77], *Logistic Regression* (LR) classifiers [28], *k-Nearest Neighbors* (NN) classifiers [32], and *Support Vector Machines* (SVM), which though simple and intuitive have proved to be one of the most effective estimation techniques [115]. A recent comparison of ML methods for binary classification is found in [4].

SVM is a supervised ML algorithm tracing back to the seminal contribution of [103], which has received significant attention in the optimization literature and has strong orientation towards real-world applications [64]. Given a set of training observations, each labeled as belonging to one of two classes, SVM goal is to detect a hyperplane induced from the available examples that is able to predict the category of new unlabeled observations. The most basic version of the SVM is the *Hard Margin*-SVM (HM-SVM) that assumes that there exists a hyperplane geometrically separating data points into the two classes, such that no observation is misclassified and margins are maximized. When the data is linearly inseparable, the *Soft Margin* SVM (SM-SVM) introduces slack variables into the constraints and aims at finding a separating hyperplane that not only achieves the maximum margin between the two classes but also minimizes the training error of misclassification [8, 25]. Many variations to the classical SVM approach have been proposed over time to enhance the predictive power of classifiers, see for instance [16, 49, 55, 59, 67, 79, 104].

In this paper, we specifically focus our attention on the SVM variant presented in [59], whose computational experience proved to detect separators with higher levels of accuracy compared to the standard ones. This method, rather than directly trying to minimize the classification error with respect to a single hyperplane, suggests to separate the sets by firstly finding two parallel hyperplanes so that the intersection of the convex hulls of the two sets is contained between them, and then to construct a third hyperplane parallel to and laying between the previous ones and such that the number of misclassified observations is minimized.

An underlying assumption of classical SVM approaches is that the input observations are not corrupted with noise and, therefore, all problem data are known exactly at the moment of classifying [21]. This assumption, however, is not always practical. Indeed, real-world observations are often plagued by uncertainty (*e.g.*, due to limited precision of collecting instruments, measurement mistakes in data gathering, sampling errors, etc.) and disregarding it might lead to solutions that are far from optimal, as well as to major fluctuations of performances [41]. Therefore, the problem of designing classifiers not facing deterioration when there are some perturbations in the data set is an interesting problem that has gained considerable attention from the scientific community. One of the main paradigms to deal with problems affected with uncertain data is given by *Robust Optimization* (RO) (see [7] and [9]). RO addresses the uncertain nature of a problem without making any specific assumption on the probability distribution of the underlying uncertain parameter, which is instead assumed to belong to a prespecified uncertainty set. RO then adopts a min-max approach that addresses uncertainty by guaranteeing the feasibility and optimality of the solution against all instances of the parameter within the uncertainty set region. Another way to handle uncertainty is given by *Distributionally Robust Optimization* (DRO) pioneered in [78] and [117], which can be regarded as a natural generalization of *Stochastic Programming* (SP) and RO. In DRO optimal decisions are sought for the worst-case probability distribution within a family of possible distributions defined by certain properties. The two most widely used types of ambiguity sets in the DRO literature are moment-based and statistical distance-based sets. While moment-based ambiguity sets contain all probability distributions that satisfy certain general moment conditions, the statistical distance-based approach considers distributions that are close in the sense of a chosen distance to a nominal distribution (*e.g.*, the empirical one). Popular choices to measure the dissimilarity between two probability distributions are Wasserstein distance or $\phi$-divergences (such as Kullback-Leibler divergence, Burg entropy, Modified-$\chi^2$ distance, Variation distance, etc.). A growing literature in these directions both from theoretical and applied points of view can be found in [1, 5, 27, 38, 40, 83, 84, 110, 116, 120].

In this paper, we deal with the binary classification problem under feature uncertainty of the input data, introducing robust and distributionally robust versions of one of the deterministic formulations presented in [59] (Formulation II), aiming at obtaining a classifier that has good generalization properties and reduces the error of misclassification of new unseen data. The main contributions of the paper are four-fold and can be summarized as follows:

- To develop box and ellipsoidal robust counterparts of the deterministic model associated with the Formulation II proposed in [59]. We assume each input observation to be bounded within hyperrectangles and hyperellipsoids.

- To formulate a new moment-based distributionally robust counterpart associated with the Formulation II proposed in [59]. We still assume each observation to be unknown but we mitigate the degree of conservatism enforcing limits on the deviations along directions detected by means of *Principal Component Analysis* (PCA) [46].

- To provide extensive numerical experiments based on real-world databases [29] with the aim of understanding the advantage of explicitly considering the uncertainty versus deterministic approaches.

- To provide managerial insights on how to choose between robust and distributionally robust approaches to model uncertainty, depending on the data set dimension.

The paper is organized as follows. Section 2 provides a literature review, while Section 3 presents basic facts and notation. In Section 4 we introduce new robust and distributionally robust optimization models for SVM, along with tractable reformulations. Section 5 presents experiments attempting to evaluate the accuracy of the proposed formulations versus deterministic approaches. Finally, conclusions and future works are provided in Section 6.

## 2. Literature Review

The extensive connections between RO, DRO and SVM have been explored by a number of authors. In [33] a minimax model for data bounded by hyper-rectangles is presented. The model looks for a linear hyperplane that minimizes the worst-case loss over input data in given intervals, and a tractable reformulation in the form of *Linear Programming* (LP) is provided. In [13, 14, 15] *Second Order Cone Programming* (SOCP) formulations are derived to design linear classifiers when the uncertainty of input observations is described by multivariate normal distributions. Geometrically, these solutions correspond to a minimax strategy with hyper-ellipsoids around the training instances, rather than hyper-rectangles. Similar approaches are provided in [97, 98], where the additive perturbations of the uncertain data are assumed to be bounded by the general $w$-norm. A related model is [16] that, assuming the data to be subject to additive noises bounded by the general $w$-norm, constructs classifiers by focusing on the more trust-worthy data that are less uncertain. A more general case for bounded uncertainty sets is studied in [113], where the linkage between regularization and robustness is also showed. The authors proved that, even though traditional SVM methods do not explicitly consider individual data uncertainties, the objective function regularization term aimed at maximizing the classifier margins represents a kind of intrinsic robustness. Other important insights about stability of SVM against uncertainty with bounded sets are due to [96], while the work developed in [48, 72] demonstrate how robust classification can be used to handle situations with imbalanced training data. For other models with polyhedral uncertainty sets see [34, 36]. Detailed reviews of the existing literature on RO in ML can be found in [20].

RO and DRO are also used for solving *Chance-Constrained* (CC)-SVM, to ensure bounded probabilities of misclassification for the uncertain data. In [52] the authors consider the case of binary classification, where only the mean and covariance matrix of the classes are assumed to be known. The minimax probabilistic decision hyperplane is then determined by optimizing the worst-case probabilities over all possible class-conditional distributions. Besides, the model presented in [86] treats all input observations as random variables for which only finite mean and covariance matrices are known, and then looks for the hyperplane able to correctly classify the observations, with high probability, even for the worst distributions. Both of these CC-SVM are relaxed using Chebyshev inequality ([66]) to yield a SOCP whose solution is guaranteed to satisfy the original problem. In a similar fashion, the Bernstein bounding scheme ([75]) is used in [6, 12]. Under the same assumptions of known moments, equivalent results have been obtained in [107], where the authors propose a different proof for obtaining the equivalent SOCP formulation and also provide reformulations in the form of *Semidefinite Programming* (SDP) models. Analogously, *Pearson divergence* distributionally robust CC-SVM is discussed in [85]. Another related work is [106], which investigates the stochastic sub-gradient descent method to solve distributionally robust CC-SVM on large-scale data sets.

In [43] risk averse theory is linked to SVM, showing that the minimization of a convex risk functional in place of the traditional hinge-loss objective function (*i.e.*, minimization of the empirical risk) straightforwardly treats a class of DRO problems. This corresponds to build an ambiguity set for the population distribution based on samples, and then searching for the classifier that minimizes the sum of the regularization term and the hinge-loss function for the worst-case distribution within the set. Authors also prove that under a specific class of risk functionals the distributionally robustified models can be reformulated as tractable convex optimization problems. Risk averse SVM is further investigated in [105] where the authors, instead of using a single measure of risk as SVM objective function, propose group differentiation by employing a different risk functional for every single class. Other related studies are [44, 93, 99, 114]. In a similar fashion, DRO for classification problems with Wasserstein ambiguity set has been investigated in [50, 54]. Instead of solving an optimization problem minimizing the hinge-losses of misclassified samples, the proposed formulation minimizes the worst-case expected prediction error with respect to distributions belonging to a Kantorovich ball, which is centered on the empirical distribution based on samples. Related works are [57, 63]. Learning and classification algorithms have also been proposed under the $\phi$-divergence measures, see for instance [30, 31], and with ambiguity sets measured via maximum mean discrepancy, see [90]. All these approaches for linear SVM models are summarized in Table 1.

|  | *Uncertainty* |  |  | *Methodology* |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | *Features Uncertainty* | *Labels Uncertainty* | *Missing Data Uncertainty* | *Box RO* | *Ellipsoidal RO* | *Bounded by norm RO* | *Polyhedral RO* | *Chance Constraints* | *Moments DRO* | *Wasserstein DRO* | *$\phi$-divergences DRO* | *Mean Discrepancy DRO* | *Risk Averse* |
| Lanckriet et al. (2002), [52] | ✓ |  |  |  |  |  |  | ✓ | ✓ |  |  |  |  |
| El Ghaoui et al. (2003), [33] | ✓ | ✓ |  | ✓ |  |  |  |  |  |  |  |  |  |
| Fung et al. (2003), [36] | ✓ |  |  |  |  |  | ✓ |  |  |  |  |  |  |
| Bhattacharyya et al. (2004, 2005), [13, 14, 15] | ✓ |  | ✓ |  | ✓ |  |  |  |  |  |  |  |  |
| Bi and Zhang (2005), [16] | ✓ |  |  |  |  | ✓ |  |  |  |  |  |  |  |
| Shivaswamy et al. (2006), [86] | ✓ |  | ✓ |  |  |  |  | ✓ | ✓ |  |  |  |  |
| Trafalis and Gilbert (2006, 2007), [97, 98] | ✓ |  |  |  |  | ✓ |  |  |  |  |  |  |  |
| Takeda and Kanamori (2009), [93] | ✓ |  |  |  |  |  |  |  |  |  |  |  | ✓ |
| Bhadra et al. (2009), [12] | ✓ |  |  |  |  |  |  | ✓ | ✓ |  |  |  |  |
| Xu et al. (2009), [113] | ✓ |  |  |  |  | ✓ |  |  |  |  |  |  |  |
| Trafalis and Alwazzi (2010), [96] | ✓ |  |  |  |  | ✓ |  |  |  |  |  |  |  |
| Ben-Tal et al. (2011), [6] | ✓ |  |  |  |  |  |  | ✓ | ✓ |  |  |  |  |
| Pant et al. (2011), [72] | ✓ |  |  |  |  | ✓ |  |  |  |  |  |  |  |
| Tsyurmasto et al. (2013), [99] | ✓ |  |  |  |  |  |  |  |  |  |  |  | ✓ |
| Fan et al. (2014), [34] | ✓ |  |  |  |  |  | ✓ |  |  |  |  |  |  |
| Gotoh et al. (2014), [44] | ✓ |  |  |  |  |  |  |  |  |  |  |  | ✓ |
| Katsumata and Takeda (2015), [48] | ✓ |  |  |  |  | ✓ |  |  |  |  |  |  |  |
| Lee and Mehrotra (2015), [54] | ✓ |  |  |  |  |  |  |  |  | ✓ |  |  |  |
| Gotoh and Uryasev (2017), [43] | ✓ |  |  |  |  |  |  |  |  |  |  |  | ✓ |
| Wang et al. (2017, 2018), [106, 107] | ✓ |  |  |  |  |  |  | ✓ | ✓ |  |  |  |  |
| Duchi et al. (2019, 2021), [30, 31] | ✓ |  |  |  |  |  |  |  |  |  | ✓ |  |  |
| Bertsimas et al. (2019), [10] | ✓ | ✓ |  |  |  | ✓ |  |  |  |  |  |  |  |
| Kuhn et al. (2019), [50] | ✓ |  |  |  |  |  |  |  |  | ✓ |  |  |  |
| Staib and Jegelka (2019), [90] | ✓ |  |  |  |  |  |  |  |  |  |  | ✓ |  |
| Vitt et al. (2019), [105] | ✓ |  |  |  |  |  |  |  |  |  |  |  | ✓ |
| Li at al. (2020), [57] | ✓ |  |  |  |  |  |  |  |  | ✓ |  |  |  |
| Shen at al. (2020), [85] | ✓ |  |  |  |  |  |  | ✓ |  |  | ✓ |  |  |

Table 1: Linear SVM Literature Review.

While all these approaches have dealt mainly with input data features uncertainty, there have also been attempts to model uncertainty in observation labels, see [18, 20, 68, 92, 111] and [10], where robust methods are employed to construct a new family of classifiers protecting against uncertainty in both features and labels for the three most widely used classification algorithms (*i.e.*, SVM, LR, and DT). RO is also employed in [39] to address the problem of corruption in missing data (see [37]), sensitivity to outliers in input samples ([35, 47, 53, 56, 115]) and to adversarial training [61, 80, 112, 119], where it is assumed that data become corrupted during the classification phase.

The approaches presented so far to hedge against uncertainty have also been successfully applied to many SVM variants. Robust counterparts have indeed been developed for the *Twin Support Vector Machine* (T-SVM), firstly proposed by [49]. See, for instance, [19, 62, 65, 76] and references therein. An alternative formulation, known as *$\nu$-Support Vector Machine* ($\nu$-SVM), was designed in [79], and models to hedge against uncertainty are proposed in [94, 108]. Another popular variant of SVM is the so called *One-Class Support Vector Machine* (OC-SVM) pioneered in [67], with robust reformulations that can be found in [60, 100, 101, 102]. There also has been a recent surge of interest in the ML community for developing distributionally robust SVM models aiming at fairness, which represents the need of a classifier performance to be invariant under certain sensitive perturbations of the inputs. Fairness in ML goes beyond the scope of this article, so we refer to [45, 94, 95, 109] and references therein. For a comprehensive survey of RO developments in the field of SVM we refer the reader to [88, 89].

The approach we propose in this paper substantially differs from the literature in several perspectives. Foremost, the deterministic variant we aim at robustifying is the one proposed in [59], which with the inclusion of a line search step showed to outperform the classical formulation in prediction accuracy. Besides, two streams of distributionally robust approaches have emerged from the review of SVM literature. The first poses the SVM problem as a CC program and then looks for bounding schemes that find solutions guaranteed to satisfy the probabilistic constraint in the worst-case distribution. The second stream, instead, aims at minimizing in the objective function the worst-case expected prediction error with respect to distributions belonging to a prespecified ambiguity set. Our proposal does not fall into any of these branches, since we are not dealing with CC programs or with uncertainty into the objective function, rather we consider input data to be random variables with unknown distributions, and then we optimize over the worst one affecting the coefficients of the constraints left-hand sides. Furthermore, we provide exact reformulations rather than approximations.

## 3. Basic Facts and Notation

In the following, all vectors will be column vectors. We use ";" for adjoining elements in a column and "," for adjoining elements in a row. Vector components are identified as being subscripted, while superscripts specify to which observation we are referring to. Vector $e$ of arbitrary dimension has all entries equal to one, while $\mathcal{I}$ and $\mathbf{0}$ denote, respectively, the identity matrix and the square null matrix of dimension $n$. We denote by $\mathbb{R}^n$ the $n$-dimensional real space, by $\mathbb{R}^n_+$ the set of non-negative vectors of dimension $n$, by $\mathbb{N}$ the set of natural numbers and by $\text{diag}(a) \in \mathbb{R}^{n\times n}$ the matrix whose $n$ diagonal entries are the elements of vector $a$ and off-diagonal components are all equal to zero. For any vector $a \in \mathbb{R}^n$, $|a| \in \mathbb{R}^n_+$ represents the vector of absolute values of the components of $a$, *i.e.*, $|a| := [|a_1|; |a_2|; \dots ; |a_n|]$. For any vector $a \in \mathbb{R}^n$ and $1 \leq w < \infty$, its $w$-norm is defined as $\|a\|_w$ with:

$$\|a\|_w := \left(\sum_{p=1}^n |a_p|^w\right)^{\frac{1}{w}} \quad \text{and} \quad \|a\|_\infty := \max_{p=1,\dots,n} |a_p|.$$

Finally, the indicator function $\mathbb{1}(\alpha \in \mathbb{R}) = 1$ if $\alpha > 0$, and 0 otherwise.

### *3.1. The Classification Problem*

Let $X$ and $Y$ be two sets of points such that $X := \left\{x^{(1)}, x^{(2)}, \dots, x^{(I)}\right\} \subseteq \mathbb{R}^n$ and $Y := \left\{y^{(1)}, y^{(2)}, \dots, y^{(J)}\right\} \subseteq \mathbb{R}^n$.

The *Hard Margin* SVM (HM-SVM) separating hyperplane is defined by a pair $(a \in \mathbb{R}^n, \gamma \in \mathbb{R})$ such that all vectors in $X$ lie on one side of the hyperplane, all the vectors in $Y$ lie on the opposite side and the distance between the separating hyperplane and the nearest data point of each class is maximized [103]. The HM-SVM optimization problem is defined as follows:

$$\begin{aligned} \min_{a,\gamma} \quad & \|a\|_w \\ \text{s.t.} \quad & a^\top x^{(i)} \leq \gamma - 1 && i = 1, \dots, I \\ & a^\top y^{(j)} \geq \gamma + 1 && j = 1, \dots, J, \end{aligned} \tag{1}$$

whose solution maximizes the distance between the hyperplanes $(a, \gamma - 1)$ and $(a, \gamma + 1)$ computed using the dual norm $\|\cdot\|_v$ with $\frac{1}{v} + \frac{1}{w} = 1$. The dual norm of the 1-norm is the infinity norm, and vice versa.

*Soft Margin* SVM (SM-SVM) relaxes the condition of perfect separability, introducing slack variables in the constraints and penalizing in the objective function data points belonging to the wrong side of the hyperplane. Specifically, let $z_X := [z_{x^{(1)}}; \dots ; z_{x^{(I)}}] \in \mathbb{R}^I_+$ and $z_Y := [z_{y^{(1)}}; \dots ; z_{y^{(J)}}] \in \mathbb{R}^J_+$ be the non-negative vectors of errors of group $X$ and $Y$. Observation $x^{(i)} \in \mathbb{R}^n$ will be correctly classified if $0 \leq z_{x^{(i)}} \leq 1$, or misclassified if $z_{x^{(i)}} > 1$. Similarly, for every observation $y^{(j)} \in \mathbb{R}^n$. The SM-SVM optimization problem is then defined as follows [25]:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_w + \nu \left(e^\top z_X + e^\top z_Y\right) \\ \text{s.t.} \quad & a^\top x^{(i)} \leq \gamma - 1 + z_{x^{(i)}} && i = 1, \dots, I \\ & a^\top y^{(j)} \geq \gamma + 1 - z_{y^{(j)}} && j = 1, \dots, J \\ & z_X \geq 0,\ z_Y \geq 0, \end{aligned} \tag{2}$$

where the user-defined penalty parameter $\nu \geq 0$ is introduced to allow a trade-off between the margin maximization and tolerating misclassification.

In order to achieve superior pattern separation, rather than minimizing the classification error with respect to a single hyperplane, in [59] it is proposed to separate the sets $X$ and $Y$ by firstly finding two parallel hyperplanes $H_1$ and $H_2$ that satisfy the following properties:

**(P1)** all points of $X$ lie on one side of $H_1$;

**(P2)** all points of $Y$ lie on the opposite side of $H_2$;

**(P3)** the intersection of convex hulls of $X$ and $Y$ is contained in the region between $H_1$ and $H_2$.

Through line search, hyperplane $H_3$ is then constructed parallel to (and lying between) $H_1$ and $H_2$, such that most of the points of $X$ lie on the same side of $H_3$ and most of the points of $Y$ lie on the opposite side of $H_3$. A point that fails to do so is called a *misclassified point*. Therefore, $H_3$ should be determined so that the number of misclassified points is minimized. In [59] five different deterministic formulations are proposed for obtaining hyperplane $H_3$, and since Formulation II proves to outperform the others, we restrict our attention to it. This formulation employs as starting point the hyperplane separating algorithm detected by model (2), in which the hyperplane margins are measured by means of the $\infty$-norm, and hence requires the minimization of $\|a\|_1$ into the objective function. Once the starting hyperplane $(a, \gamma)$ of (2) is obtained, it is shifted in order to determine hyperplanes $H_1$ and $H_2$ that satisfy properties **(P1)-(P3)**. Specifically, $H_1 := (a, \gamma - 1 + \omega_1)$ and $H_2 := (a, \gamma + 1 - \omega_2)$, where:

$$\omega_1 := \max\left\{z_{x^{(i)}} \;\middle|\; i = 1, \ldots, I\right\}, \qquad \omega_2 := \max\left\{z_{y^{(j)}} \;\middle|\; j = 1, \ldots, J\right\}. \tag{3}$$

The following minimization problem is finally solved using the line search method (see [70]), with the aim of obtaining the scalar $b \in \mathbb{R}$ that defines the hyperplane $H_3 := (a, b)$, parallel to and lying between $H_1$ and $H_2$ and minimizing the overall number of misclassified points:

$$\begin{array}{ll} \min\limits_{b} & \sum\limits_{i=1}^{I} \mathbb{1}\left(a^\top x^{(i)} - b\right) + \sum\limits_{j=1}^{J} \mathbb{1}\left(b - a^\top y^{(j)}\right) \\ \text{s.t.} & \qquad \gamma + 1 - \omega_2 \leq b \leq \gamma - 1 + \omega_1. \end{array} \tag{4}$$

Specifically, as the objective of (4) is not continuous, we divide the interval $[b_{\min}, b_{\max}] := [\gamma + 1 - \omega_2, \gamma - 1 + \omega_1]$ into $k_{\max}$ sub-intervals of equal length and denote $s_k := \sum_{i=1}^{I} \mathbb{1}\left(a^\top x^{(i)} - b_k\right) + \sum_{j=1}^{J} \mathbb{1}\left(b_k - a^\top y^{(j)}\right)$, with $b_k = b_{\min} + k \cdot \frac{b_{\max} - b_{\min}}{k_{\max}}$, $k = 0, \ldots, k_{max}$. The final solution of (4) is then given by $b_{k^*}$ with $k^* \in \arg\min\{s_0, \ldots, s_k, \ldots, s_{k_{\max}}\}$.

## 4. Robust and Distributionally Robust Support Vector Machine Models

The basic assumption of the deterministic model (2)-(4) presented in [59] is that all input observations of both groups $X$ and $Y$ are always provided exactly, ignoring any type of uncertainty associated with lack of data or with data that cannot be fully trusted. However, when the given values differ significantly from the true ones, the predictive power of the deterministic classifier might be unsatisfactory. Therefore in this section, rather than dealing with a countable set of well-defined data points, we handle data features as uncertain and formulate robust counterparts to model problem (2)-(4) with uncertainty sets in the form of hyperrectangles (Section 4.1.1) and hyperellipsoids (Section 4.1.2). Moreover, we propose a distributionally robust counterpart to the deterministic formulation (2)-(4) that enforces limits on the observations first-order deviations along directions detected by means of PCA (Section 4.2).

### *4.1. Robust Support Vector Machine*

In this section, we assume the uncertainty of every input observation $x^{(i)} \in X \subseteq \mathbb{R}^n,\ i = 1,\dots,I$ to be represented by the uncertainty set $\mathcal{U}\big(x^{(i)}\big)$. Equivalently for every observation $y^{(j)} \in Y \subseteq \mathbb{R}^n,\ j = 1,\dots,J$. Then, the robust counterpart of model (2) that optimizes over worst-case realizations on all possible observations in $\mathcal{U}\big(x^{(i)}\big), \mathcal{U}\big(y^{(j)}\big),\ i = 1,\dots,I,\ j = 1,\dots,J$ corresponds to the following optimization model:

$$
\begin{aligned}
\min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) & & \\
\text{s.t.} \quad & \max_{x\in\mathcal{U}(x^{(i)})} \big[a^\top x\big] \leq \gamma - 1 + z_{x^{(i)}} & & i = 1,\dots,I \\
& \min_{y\in\mathcal{U}(y^{(j)})} \big[a^\top y\big] \geq \gamma + 1 - z_{y^{(j)}} & & j = 1,\dots,J \\
& z_X \geq 0,\ z_Y \geq 0. & &
\end{aligned} \tag{5}
$$

The size of the uncertainty sets $\mathcal{U}\big(x^{(i)}\big), \mathcal{U}\big(y^{(j)}\big),\ i = 1,\dots,I,\ j = 1,\dots,J$ reflects the degree of data uncertainty. If:

$$
\mathcal{U}\big(x^{(i)}\big) := \big\{x^{(i)}\big\},\ i = 1,\dots,I \quad \text{and} \quad \mathcal{U}\big(y^{(j)}\big) := \big\{y^{(j)}\big\},\ j = 1,\dots,J,
$$

then the robust formulation (5) reduces to the deterministic model (2).

#### *4.1.1. Robust Support Vector Machine with Interval Data Uncertainty*

First, we consider uncertainty sets having the form of hyperrectangles. Let $\zeta_{x^{(i)}}, \zeta_{y^{(j)}} \in \mathbb{R}^n_+$ define the perturbation vectors of input observations $x^{(i)}$ and $y^{(j)}$, respectively; further, let $\rho_X, \rho_Y \in \mathbb{R}_+$ be global measures of uncertainty for group $X$ and $Y$, respectively. Then, the hyperrectangular uncertainty sets $\mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big)$ and $\mathcal{U}_{\mathcal{B}}\big(y^{(j)}\big)$ centered around $x^{(i)}$ and $y^{(j)}$ are defined, respectively, as:

$$
\mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big) := \Big\{x \in \mathbb{R}^n \ \Big|\ x^{(i)} - \rho_X \zeta_{x^{(i)}} \leq x \leq x^{(i)} + \rho_X \zeta_{x^{(i)}} \Big\} \quad i = 1,\dots,I, \tag{6}
$$

$$
\mathcal{U}_{\mathcal{B}}\big(y^{(j)}\big) := \Big\{y \in \mathbb{R}^n \ \Big|\ y^{(j)} - \rho_Y \zeta_{y^{(j)}} \leq y \leq y^{(j)} + \rho_Y \zeta_{y^{(j)}} \Big\} \quad j = 1,\dots,J. \tag{7}
$$

Depending on how reliable the decision maker considers the available data, parameters $\rho_X$ and $\rho_Y$ allow to tailor the degree of conservatism. When uncertainty sets are described by means of

(6)-(7), model (5) can be reformulated by the following linear program (see [33] and derivation in Appendix A):

$$
\begin{aligned}
\min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\
\text{s.t.} \quad & a^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top |a| \leq \gamma - 1 + z_{x^{(i)}} \qquad && i = 1, \ldots, I \\
& a^\top y^{(j)} - \rho_Y \zeta_{y^{(j)}}^\top |a| \geq \gamma + 1 - z_{y^{(j)}} \qquad && j = 1, \ldots, J \\
& z_X \geq 0, \ z_Y \geq 0,
\end{aligned} \tag{8}
$$

where the number of continuous variables is $n+1+I+J$ and the number of constraints is $2(I+J)$ of which $I+J$ are non-negative. As in the deterministic case, once the solution $(a, \gamma, z_X, z_Y)$ of (8) is obtained, the final hyperplane $H_3$ is recovered through line search:

$$
\begin{aligned}
\min_{b} \quad & \sum_{i=1}^{I} \mathbb{1}\big(a^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top |a| - b\big) + \sum_{j=1}^{J} \mathbb{1}\big(b - a^\top y^{(j)} + \rho_Y \zeta_{y^{(j)}}^\top |a|\big) \\
\text{s.t.} \quad & \gamma + 1 - \omega_2 \leq b \leq \gamma - 1 + \omega_1,
\end{aligned} \tag{9}
$$

with $\omega_1, \omega_2$ as in (3), $\mathbb{1}$ is the unit step function and where robustness fails for those points whose hyperrectangle intersects the hyperplane $H_3$. Consequently, all those points either lying on the wrong side of $(a, b)$ or whose hyperrectangles intersect $H_3$ will be considered misclassified. To summarize, the geometrical interpretation of the proposed approach is sketched in Figure 1. For the sake of clarity, we restrict our attention to the bidimensional case ($n = 2$). Points of group $X$ are represented by filled black dots, while points of group $Y$ by empty white circles.

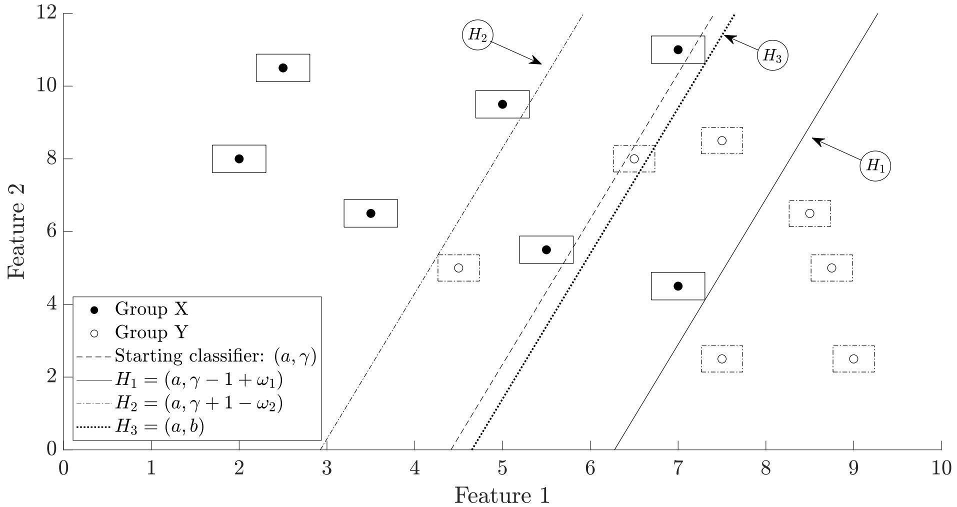


Figure 1: Input observations of groups $X$ and $Y$ bounded by boxes and separating hyperplanes $H_1$, $H_2$ and $H_3$.

After building boxes around every observation, we detect the starting hyperplane $(a, \gamma)$ by means of model (8). We then shift it to the right and to the left, by amounts $\omega_1$ and $\omega_2$ respectively, to detect $H_1$ and $H_2$ such that all boxes of group $X$ lie on one side of $H_1$ and all boxes of group $Y$ lie on the opposite side of $H_2$. Through model (9), the final classifier $H_3$ is found such that the overall number of misclassified boxes is minimized.

### *4.1.2. Robust Support Vector Machine with Ellipsoidal Data Uncertainty*

It is well known that intervals perturbations assumption can lead to overly conservative solutions. Therefore, to alleviate this drawback, in this section we propose an alternative robust formulation that considers uncertainty sets having the form of hyperellipsoids. This latter choice turns into less conservative models with respect to the hyperrectangles case since it disregards those situations under which all features jointly assume extreme interval values. Moreover, this choice does not hinder the tractability of the associated reformulation, leading to a conic quadratic program.

Let $\Sigma_{x^{(i)}}, \Sigma_{y^{(j)}} \in \mathbb{R}^{n\times n}$ be positive definite covariance matrices associated to points $x^{(i)}$ and $y^{(j)}$, respectively; further, let $\rho_X, \rho_Y \in \mathbb{R}_+$ denote the radii of the ellipsoids of groups $X$ and $Y$, respectively. Then, the ellipsoidal uncertainty sets $\mathcal{U}_{\mathcal{E}}\big(x^{(i)}\big)$ and $\mathcal{U}_{\mathcal{E}}\big(y^{(j)}\big)$ centered around $x^{(i)}$ and $y^{(j)}$ are defined, respectively, as:

$$\mathcal{U}_{\mathcal{E}}\big(x^{(i)}\big) := \left\{ x \in \mathbb{R}^n \;\middle|\; \big(x - x^{(i)}\big)^\top \Sigma^{-1}_{x^{(i)}} \big(x - x^{(i)}\big) \le \rho_X^2 \right\} \quad i = 1, \dots, I, \tag{10}$$

$$\mathcal{U}_{\mathcal{E}}\big(y^{(j)}\big) := \left\{ y \in \mathbb{R}^n \;\middle|\; \big(y - y^{(j)}\big)^\top \Sigma^{-1}_{y^{(j)}} \big(y - y^{(j)}\big) \le \rho_Y^2 \right\} \quad j = 1, \dots, J. \tag{11}$$

According to [13] (see derivation in Appendix A), when uncertainty sets are described by means of (10)-(11), model (5) can be reformulated by the following SOCP:

$$\begin{aligned}
\min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\
\text{s.t.} \quad & a^\top x^{(i)} + \rho_X \|\Sigma^{\frac{1}{2}}_{x^{(i)}} a\|_2 \le \gamma - 1 + z_{x^{(i)}} && i = 1, \dots, I \\
& a^\top y^{(j)} - \rho_Y \|\Sigma^{\frac{1}{2}}_{y^{(j)}} a\|_2 \ge \gamma + 1 - z_{y^{(j)}} && j = 1, \dots, J \\
& z_X \ge 0, \; z_Y \ge 0,
\end{aligned} \tag{12}$$

where:

$$\|\Sigma^{\frac{1}{2}}_{x^{(i)}} a\|_2 := \sqrt{a^\top \Sigma_{x^{(i)}} a} \quad \text{and} \quad \|\Sigma^{\frac{1}{2}}_{y^{(j)}} a\|_2 := \sqrt{a^\top \Sigma_{y^{(j)}} a}.$$

The number of continuous variables is $n + 1 + I + J$, while the number of constraints is $2(I + J)$ of which $I + J$ are non-negative. From the solution of problem (12), we get hyperplanes $H_1$ and $H_2$ which satisfy properties **(P1)-(P3)** with $\omega_1, \omega_2$ as in (3). To find $H_3$ we finally solve the following minimization problem using line search:

$$\begin{aligned} \min_{b} \quad & \sum_{i=1}^{I} \mathbb{1}\big(a^\top x^{(i)} + \rho_X \|\Sigma_{x^{(i)}}^{\frac{1}{2}} a\|_2 - b\big) + \sum_{j=1}^{J} \mathbb{1}\big(b - a^\top y^{(j)} + \rho_Y \|\Sigma_{y^{(j)}}^{\frac{1}{2}} a\|_2\big) \\ \text{s.t.} \quad & \gamma + 1 - \omega_2 \le b \le \gamma - 1 + \omega_1, \end{aligned} \tag{13}$$

where robustness fails for those points whose hyperellipsoid intersects the decision hyperplane $H_3$. To summarize, the geometrical interpretation of the proposed approach is sketched in Figure 2. After building ellipsoids around every observation, we detected the starting hyperplane $(a, \gamma)$ by means of model (12). We then shift it to the right and to the left, by amounts $\omega_1$ and $\omega_2$ respectively, to detect $H_1$ and $H_2$ such that all ellipsoids of group $X$ lie on one side of $H_1$ and all ellipsoids of group $Y$ lie on the opposite side of $H_2$. Through line search of model (13), the final classifier $H_3 = (a, b)$ is found such that the overall number of misclassified ellipsoids is minimized.

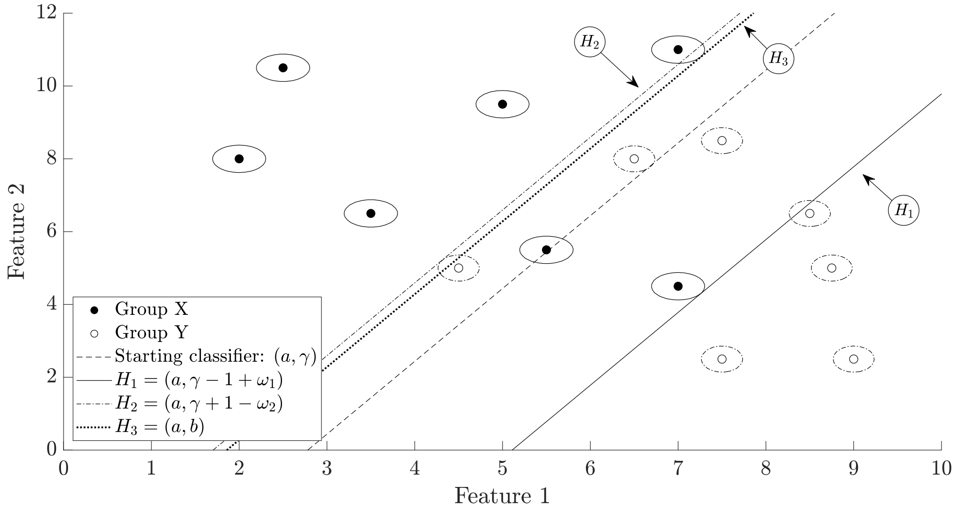


Figure 2: Input observations of groups $X$ and $Y$ bounded by ellipsoids and separating hyperplanes $H_1$, $H_2$ and $H_3$.

### *4.2. Distributionally Robust Support Vector Machine*

Solutions obtained considering uncertainty sets having the form of hyperellipsoids can still be too conservative. One way to overcome this limitation would consist in resorting to other types of uncertainty sets such as polyhedral, conic, convex constraints (see [42]) or combinations of them (*e.g.*, box + ellipsoidal, box + polyhedral, box + ellipsoidal + polyhedral, see [58]). However, these specific approaches would require precise knowledge of the instances under analysis and would be highly problem-dependent. Moreover, conic uncertainty sets would require the use of conic duality while convex constraints sets the use of Fenchel duality.

Therefore, with the aim of providing progressively less conservative models that do not lose generalization ability and still protect against uncertainty, in this section we employ the most recent techniques of moment-based DRO.

In this section we treat all input observations $x^{(i)}$, $y^{(j)}$, $i = 1, \dots, I$, $j = 1, \dots, J$ as random variables, for which the exact probability distributions $\mathbb{P}^{\text{true}}_{x^{(i)}}$, $i = 1, \dots, I$ and $\mathbb{P}^{\text{true}}_{y^{(j)}}$, $j = 1, \dots, J$ are unknown. To hedge against uncertainty, for each input observation $x^{(i)}$ we optimize against the worst-case expectation under all possible distributions $\mathbb{P}$ belonging to the ambiguity set $\mathcal{D}\big(x^{(i)}\big)$. Equivalently for $y^{(j)}$ and $\mathcal{D}\big(y^{(j)}\big)$. Accordingly, the distributionally robust counterpart of model (2) can be formulated as follows:

$$\begin{aligned}
\min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\
\text{s.t.} \quad & \sup_{\mathbb{P}\in\mathcal{D}(x^{(i)})} \mathbb{E}_{\mathbb{P}}\Big[a^\top x\Big] \leq \gamma - 1 + z_{x^{(i)}} && i = 1, \dots, I \\
& \inf_{\mathbb{P}\in\mathcal{D}(y^{(j)})} \mathbb{E}_{\mathbb{P}}\Big[a^\top y\Big] \geq \gamma + 1 - z_{y^{(j)}} && j = 1, \dots, J \\
& z_X \geq 0,\ z_Y \geq 0.
\end{aligned} \tag{14}$$

The choice of the specific ambiguity set $\mathcal{D}$ when modeling a problem is context dependent. This decision depends on the data being represented by the set, as well as the needs of the modeler. Hereby, a formulation that protects against uncertainty not losing generalization ability is sought and we assume that estimates are easily available from a prior statistical analysis of the uncertain data. In the following, namely, we will focus on principal directions and variance information since shared by many different distributions, while disregarding higher order moments which are often unavailable (see [71]).

We consider the general class moment-based ambiguity set proposed in [110] where the support and a list of partial moments describing the uncertainty are available:

$$\mathcal{D}\left(x^{(i)}\right) := \left\{ \mathbb{P} \in \mathcal{P}^n_+ \;\middle|\; \begin{array}{c} \mathbb{P}\Big(x \in \mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big)\Big) = 1 \\ \mathbb{E}_{\mathbb{P}}\big[g_p(x)\big] \leq \big(\varrho_X\big)_p \quad p = 1, \dots, n \end{array} \right\} \quad i = 1, \dots, I, \tag{15}$$

with $\mathcal{P}^n_+$ representing the set of probabilities distributions on $\mathbb{R}^n$. Specifically, the first constraint in set (15) requires every realization to be constrained within its support set $\mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big)$ defined in (6). The second group of constraints in (15) characterizes the moments information via $n$ functions $g_p(\cdot)$, and enforces the generalized moment $\mathbb{E}_{\mathbb{P}}\big[g_p(x)\big]$ not to exceed a given threshold $\big(\varrho_X\big)_p \in \mathbb{R}_+$, $p = 1, \dots, n$. While several generalized moment functions describing moment information were suggested in the literature, in this paper we employ the piecewise linear formulation proposed by [3], which can be interpreted as the first-order deviations of the uncertain parameter with respect to the nominal value $x^{(i)}$ along certain projections $f^{(p)}_X \in \mathbb{R}^n$. Namely:

$$g_p(x) := \left| {f^{(p)}_X}^\top \left(x - x^{(i)}\right) \right| \quad p = 1, \dots, n. \tag{16}$$

To determine projections $F_X := \big[f_X^{(1)}, \dots, f_X^{(n)}\big] \in \mathbb{R}^{n\times n}$ and thresholds $\varrho_X := \big[(\varrho_X)_1; \dots; (\varrho_X)_n\big] \in \mathbb{R}_+^n$ we adopt a database strategy based on PCA (see [82]). The same approach holds for observations of group $Y$.

1. Given an unbiased estimate of the covariance matrix $\Sigma_X$:

$$\Sigma_X := \frac{\left(\sum\limits_{i=1}^{I} {x^{(i)}}^\top x^{(i)}\right) - \left(\sum\limits_{i=1}^{I} x^{(i)}\right)^\top \left(\sum\limits_{i=1}^{I} x^{(i)}\right)}{I-1}, \tag{17}$$

   we perform PCA onto $\Sigma_X$. Performing PCA enables capturing meaningful information about the available data. Specifically, it enables detecting the directions that manifest the most variations. We obtain:

$$\Sigma_X = F_X \cdot \Lambda_X \cdot F_X^\top, \tag{18}$$

   where $F_X \in \mathbb{R}^{n\times n}$ stands for the orthogonal transformation matrix and $\Lambda_X := \mathrm{diag}(\lambda_X) \in \mathbb{R}_+^{n\times n}$ is a diagonal matrix including variance information $\lambda_X$ after transformation (*i.e.*, along the principal directions $F_X$).

2. To determine the maximum deviations allowed along the $n$ principal directions given by thresholds $\varrho_X := \big[(\varrho_X)_1; \dots; (\varrho_X)_n\big] \in \mathbb{R}_+^n$ we set:

$$(\varrho_X)_p := \frac{\rho_X \sqrt{(\lambda_X)_p}}{K} \qquad p = 1, \dots, n, \tag{19}$$

   where $\lambda_X$ has been obtained from PCA and $K \in \mathbb{N} \setminus \{0\}$ is a scale parameter.

An attractive feature of this moment-based approach, is that one can control the model degree of conservatism simply by adjusting values of the limits $(\varrho_X)_p$, $p = 1, \dots, n$. So, depending on specific applications and problem instances, one can opt for a more conservative strategy and tune lower values for the scale parameter $K$, or opt for more aggressive approaches setting higher values of $K$ and allowing less dispersion.

Figure 3 provides a graphical representation of the procedure for a single observation $x^{(i)}$. Given a starting group $X$, PCA is performed to detect principal directions $F_X = \big[f_X^{(1)}, f_X^{(2)}\big] \in \mathbb{R}^{2\times 2}$. Then, for every observation $x^{(i)}$ the box support $\mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big)$ is defined and limits $\varrho_X = \big[(\varrho_X)_1; (\varrho_X)_2\big] \in \mathbb{R}_+^2$ on variations along principal direction $f_X^{(1)}$ and direction $f_X^{(2)}$ are enforced. As shown, tuning higher values for the scale parameter $K$ turns into a less conservative strategy compared to lower values of $K$, as allowing less dispersion.

Notice that, there are no theoretical guarantees to ensure that any moment-based ambiguity set contains the true distribution with high probability. To remedy this situation [27] proposed confidence regions for the mean and covariance matrix of the uncertainty using historical samples. Recently, other methods on data-driven robust optimization have also been suggested (see for

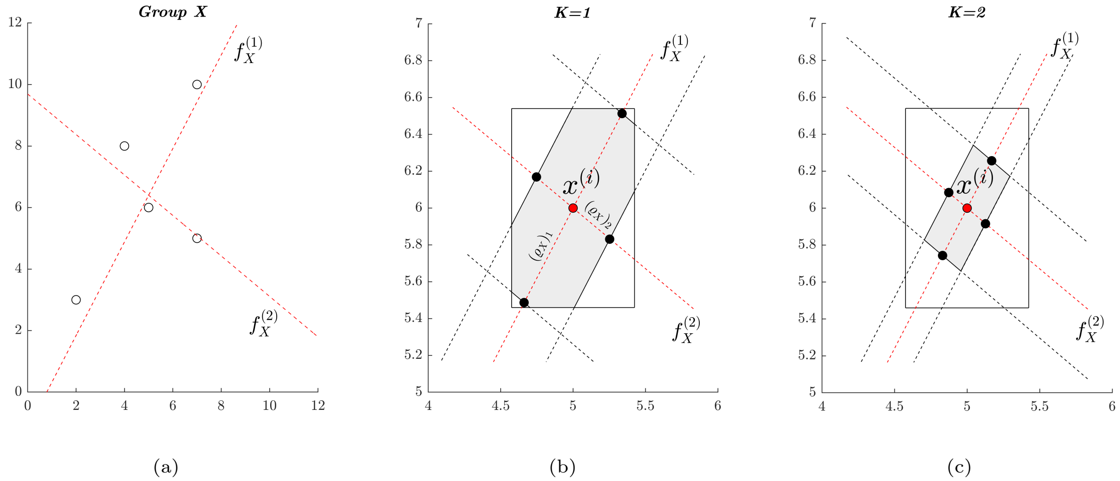


Figure 3: Given group $X$ (a), principal directions $f_X^{(1)}$ and $f_X^{(2)}$ are detected. For every point $x^{(i)}$, limits $(\varrho_X)_1$, $(\varrho_X)_2$ on variations along them are enforced together with the box support; $K$ may be fixed equal to 1 (b) or 2 (c).

instance [11], [69], [81]) employing hypothesis tests to determine the size of the ambiguity sets in order to ensure them to be statistically interpretable. Confidence regions can also be constructed from historical observations using resampling techniques, such as jackknifing or bootstrapping [24]. Unfortunately, these strategies cannot be trivially applied to the ambiguity set used in this work but represent an interesting future research direction to obtain a probabilistic guarantee for the true distribution to be contained in $\mathcal{D}$.

### *4.2.1. Tractable Reformulation*

Model (14) is intractable due to the infinite number of probability distributions contained in every ambiguity set (15); therefore, in this section, we reformulate this problem as a tractable deterministic optimization model. Introducing the auxiliary random vector $\varphi_X := [(\varphi_X)_1; \dots; (\varphi_X)_n] \in \mathbb{R}_+^n$ the ambiguity set given in (15) can be equivalently re-formulated as the projection of an extended ambiguity set $\bar{\mathcal{D}}\left(x^{(i)}\right)$:

$$\bar{\mathcal{D}}\left(x^{(i)}\right) := \left\{ \mathbb{Q} \in \mathcal{P}_+^n \;\middle|\; \begin{array}{c} \mathbb{Q}\Big(x, \varphi_X \in \bar{\mathcal{U}}_{\mathcal{B}}\big(x^{(i)}\big)\Big) = 1 \\ \mathbb{E}_{\mathbb{Q}}\left[(\varphi_X)_p\right] \le \left(\varrho_X\right)_p \quad p = 1, \dots, n \end{array} \right\} \quad i = 1, \dots, I, \tag{20}$$

with lifted support set defined as:

$$\bar{\mathcal{U}}_{\mathcal{B}}\big(x^{(i)}\big) := \left\{ \Big(x, \varphi_X\Big) \in \mathbb{R}^n \times \mathbb{R}_+^n \;\middle|\; \begin{array}{c} x \in \mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big) \\ g_p(x) \le (\varphi_X)_p \quad p = 1, \dots, n \end{array} \right\} \quad i = 1, \dots, I. \tag{21}$$

Using (6) and (16) the lifted support set (21) can be equally expressed as:

$$\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)}) = \left\{ \Big(x, \varphi_X\Big) \;\middle|\; \begin{array}{c} x \le x^{(i)} + \rho_X \zeta_{x^{(i)}} \\ x \ge x^{(i)} - \rho_X \zeta_{x^{(i)}} \\ (\varphi_X)_p \ge 0 \quad p = 1, \ldots, n \\ {f_X^{(p)}}^\top x - {f_X^{(p)}}^\top x^{(i)} \le (\varphi_X)_p \quad p = 1, \ldots, n \\ {f_X^{(p)}}^\top x^{(i)} - {f_X^{(p)}}^\top x \le (\varphi_X)_p \quad p = 1, \ldots, n \end{array} \right\} \quad i = 1, \ldots, I, \tag{22}$$

or equivalently in matrix form:

$$\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)}) = \left\{ \Big(x, \varphi_X\Big) \;\Big|\; C_X x + D\varphi_X \le h_{x^{(i)}} \right\} \quad i = 1, \ldots, I, \tag{23}$$

where:

$$C_X := \begin{bmatrix} \mathcal{I} \\ -\mathcal{I} \\ \mathbf{0} \\ F_X^\top \\ -F_X^\top \end{bmatrix} \in \mathbb{R}^{5n \times n}, \qquad D := \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ -\mathcal{I} \\ -\mathcal{I} \\ -\mathcal{I} \end{bmatrix} \in \mathbb{R}^{5n \times n}, \qquad h_{x^{(i)}} := \begin{bmatrix} x^{(i)} + \rho_X \zeta_{x^{(i)}} \\ -x^{(i)} + \rho_X \zeta_{x^{(i)}} \\ \mathbf{0} \\ F_X^\top x^{(i)} \\ -F_X^\top x^{(i)} \end{bmatrix} \in \mathbb{R}^{5n}.$$

An analogous reformulation can be performed for every observation of group $Y$ and a distributionally robust formulation equivalent to (14) is then given as follows:

$$\begin{array}{lll} \min\limits_{a, \gamma, z_X, z_Y} & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) & \\ \text{s.t.} & \sup\limits_{\mathbb{Q} \in \bar{\mathcal{D}}(x^{(i)})} \mathbb{E}_{\mathbb{Q}}\Big[a^\top x\Big] \le \gamma - 1 + z_{x^{(i)}} & i = 1, \ldots, I \\ & \inf\limits_{\mathbb{Q} \in \bar{\mathcal{D}}(y^{(j)})} \mathbb{E}_{\mathbb{Q}}\Big[a^\top y\Big] \ge \gamma + 1 - z_{y^{(j)}} & j = 1, \ldots, J \\ & z_X \ge 0, \; z_Y \ge 0. & \end{array} \tag{24}$$

It is worth noticing that the second group of constraints of formulation (24) can be expressed as:

$$\inf_{\mathbb{Q} \in \bar{\mathcal{D}}(y^{(j)})} \mathbb{E}_{\mathbb{Q}}\Big[a^\top y\Big] \ge \gamma + 1 - z_{y^{(j)}} \quad \Leftrightarrow \quad \sup_{\mathbb{Q} \in \bar{\mathcal{D}}(y^{(j)})} \mathbb{E}_{\mathbb{Q}}\Big[-a^\top y\Big] \le -\gamma - 1 + z_{y^{(j)}} \quad j = 1, \ldots, J.$$

For every $i = 1, \ldots, I$, the left-hand side of the distributionally robust constraint of model (24) coincides with the optimal value of the following moment problem:

$$\begin{aligned}\sup_{\mathbb{Q}\in\bar{\mathcal{D}}(x^{(i)})} \mathbb{E}_{\mathbb{Q}}\Big[a^\top x\Big] = \quad \sup_{\mathbb{Q}} \quad & \int_{\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)})} q(x,\varphi_X)\Big(a^\top x\Big)\,\mathrm{d}x\,\mathrm{d}\varphi_X \\ \text{s.t.} \quad & \int_{\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)})} q(x,\varphi_X)\,\mathrm{d}x\,\mathrm{d}\varphi_X = 1 \\ & \int_{\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)})} q(x,\varphi_X)\varphi_X\,\mathrm{d}x\,\mathrm{d}\varphi_X \le \varrho_X,\end{aligned} \tag{25}$$

where the decision variable is $q(x,\varphi_X)$. Introducing the multipliers $\eta_{x^{(i)}} \in \mathbb{R}$ and $\beta_{x^{(i)}} \in \mathbb{R}^n_+$, the Lagrangian reformulation of (25) is:

$$\sup_{\mathbb{Q}} \quad \int_{\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)})} q(x,\varphi_X)\left(a^\top x - \eta_{x^{(i)}} - \beta_{x^{(i)}}^\top\varphi_X\right)\,\mathrm{d}x\,\mathrm{d}\varphi_X + \eta_{x^{(i)}} + \beta_{x^{(i)}}^\top\varrho_X. \tag{26}$$

If there exists $(x,\varphi_X)$ such that $a^\top x - \beta_{x^{(i)}}^\top\varphi_X \ge \eta_{x^{(i)}}$, then (26) is unbounded above because $q(x,\varphi_X) \ge 0$, $\forall(x,\varphi_X) \in \bar{\mathcal{U}}_{\mathcal{B}}\big(x^{(i)}\big)$. Contrariwise, when $a^\top x - \beta_{x^{(i)}}^\top\varphi_X \le \eta_{x^{(i)}}$, then $\forall(x,\varphi_X) \in \bar{\mathcal{U}}_{\mathcal{B}}\big(x^{(i)}\big)$ the function admits a solution given by $\eta_{x^{(i)}} + \beta_{x^{(i)}}^\top\varrho_X$. The dual of (25) then becomes:

$$\begin{aligned}\min_{\eta_{x^{(i)}},\,\beta_{x^{(i)}}} \quad & \eta_{x^{(i)}} + \beta_{x^{(i)}}^\top\varrho_X \\ \text{s.t.} \quad & a^\top x - \beta_{x^{(i)}}^\top\varphi_X \le \eta_{x^{(i)}} \quad \forall(x,\varphi_X) \in \bar{\mathcal{U}}_{\mathcal{B}}\big(x^{(i)}\big) \\ & \beta_{x^{(i)}} \ge 0.\end{aligned} \tag{27}$$

The robust set of constraints of model (27) can be equivalently reformulated as:

$$a^\top x - \beta_{x^{(i)}}^\top\varphi_X \le \eta_{x^{(i)}} \quad \forall(x,\varphi_X) \in \bar{\mathcal{U}}_{\mathcal{B}}\big(x^{(i)}\big) \quad \Leftrightarrow \quad \max_{(x,\varphi_X)\,\in\,\bar{\mathcal{U}}_{\mathcal{B}}(x^{(i)})} \Big[a^\top x - \beta_{x^{(i)}}^\top\varphi_X\Big] \le \eta_{x^{(i)}}$$

where the dual of the left-hand side maximization problem is equal to:

$$\begin{aligned}\min_{\pi_{x^{(i)}}} \quad & \pi_{x^{(i)}}^\top h_{x^{(i)}} \\ \text{s.t.} \quad & C_X^\top\pi_{x^{(i)}} \ge a \\ & D^\top\pi_{x^{(i)}} \ge -\beta_{x^{(i)}} \\ & \pi_{x^{(i)}} \ge 0,\end{aligned} \tag{28}$$

with $\pi_{x^{(i)}} \in \mathbb{R}^{5n}_+$. Combining (27) with (28), and repeating for all $i = 1,\dots,I$ and $j = 1,\dots,J$, a tractable distributionally robust formulation of problem (2) is:

$$
\begin{array}{lll}
\min\limits_{a,\gamma,z_X,z_Y,\eta_X,\eta_Y,\beta_X,\beta_Y,\pi_X,\pi_Y} & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) & \\
\text{s.t.} & \eta_{x^{(i)}} + \beta_{x^{(i)}}^\top \varrho_X \leq \gamma - 1 + z_{x^{(i)}} & i = 1,\ldots,I \\
& \pi_{x^{(i)}}^\top h_{x^{(i)}} \leq \eta_{x^{(i)}} & i = 1,\ldots,I \\
& C_X^\top \pi_{x^{(i)}} \geq a & i = 1,\ldots,I \\
& D^\top \pi_{x^{(i)}} \geq -\beta_{x^{(i)}} & i = 1,\ldots,I \\
& \eta_{y^{(j)}} + \beta_{y^{(j)}}^\top \varrho_Y \leq -\gamma - 1 + z_{y^{(j)}} & j = 1,\ldots,J \\
& \pi_{y^{(j)}}^\top h_{y^{(j)}} \leq \eta_{y^{(j)}} & j = 1,\ldots,J \\
& C_Y^\top \pi_{y^{(j)}} \geq -a & j = 1,\ldots,J \\
& D^\top \pi_{y^{(j)}} \geq -\beta_{y^{(j)}} & j = 1,\ldots,J \\
& z_X \geq 0,\ z_Y \geq 0 & \\
& \pi_X \geq 0,\ \pi_Y \geq 0,\ \beta_X \geq 0,\ \beta_Y \geq 0, &
\end{array}
\tag{29}
$$

where $\eta_X := \left[\eta_{x^{(1)}};\ldots;\eta_{x^{(I)}}\right] \in \mathbb{R}^I$, $\eta_Y := \left[\eta_{y^{(1)}};\ldots;\eta_{y^{(J)}}\right] \in \mathbb{R}^J$, $\beta_X := \left[\beta_{x^{(1)}};\ldots;\beta_{x^{(I)}}\right] \in \mathbb{R}_+^{nI}$, $\beta_Y := \left[\beta_{y^{(1)}};\ldots;\beta_{y^{(J)}}\right] \in \mathbb{R}_+^{nJ}$, $\pi_X := \left[\pi_{x^{(1)}};\ldots;\pi_{x^{(I)}}\right] \in \mathbb{R}_+^{5nI}$ and $\pi_Y := \left[\pi_{y^{(1)}};\ldots;\pi_{y^{(J)}}\right] \in \mathbb{R}_+^{5nJ}$. The number of variables of linear formulation (29) is $n + 1 + I(2 + 6n) + J(2 + 6n)$, while the number of constraints is $n + I(5+6n) + J(5+6n)$ of which $I(1+6n) + J(1+6n)$ are non-negativity constraints. From the solution of optimization problem (29), the hyperplanes $H_1$ and $H_2$, satisfying properties **(P1)-(P3)** are obtained. We find $H_3$ solving the minimization problem via line search:

$$
\begin{array}{ll}
\min\limits_{b} & \sum\limits_{i=1}^{I} \mathbb{1}\left(\eta_{x^{(i)}} + \beta_{x^{(i)}}^\top \varrho_X - b\right) + \sum\limits_{j=1}^{J} \mathbb{1}\left(b + \eta_{y^{(j)}} + \beta_{y^{(j)}}^\top \varrho_Y\right) \\
\text{s.t.} & \gamma + 1 - \omega_2 \leq b \leq \gamma - 1 + \omega_1.
\end{array}
\tag{30}
$$

To summarize, Figure 4 provides the geometrical interpretation of the proposed approach. First, principal directions are detected for group $X$. Each nominal observation $x^{(i)}$ is therefore bounded by a box support and limits on $x^{(i)}$ deviations along principal directions are enforced. Same for every observation $y^{(j)}$ of group $Y$ and its principal directions. Then, the starting hyperplane $(a, \gamma)$ is detected by means of model (29), and it is shifted to the right and to the left by amounts $\omega_1$ and $\omega_2$, respectively, to identify $H_1$ and $H_2$. Through line search given by (30), the final classifier $H_3$ is found such that the overall number of misclassified realizations is minimized.

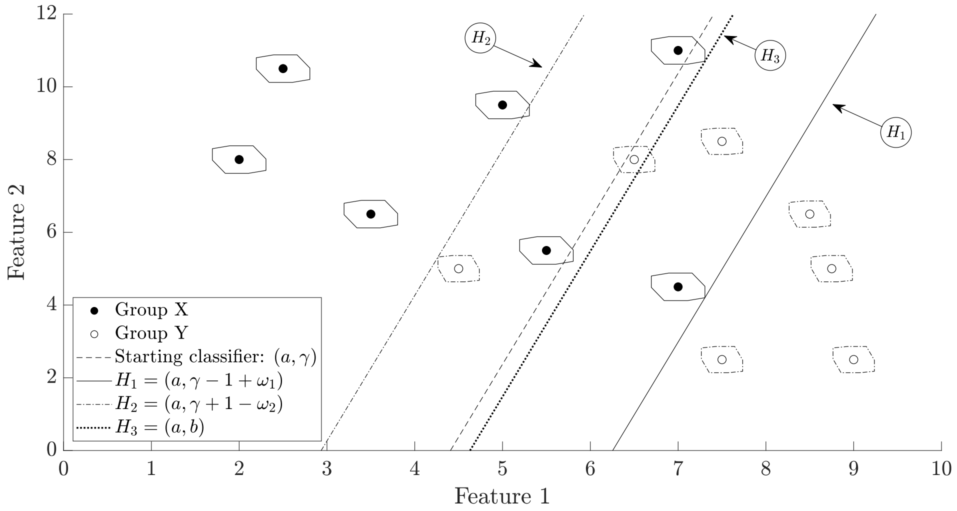


Figure 4: Input observations of groups $X$ and $Y$ and separating hyperplanes $H_1$, $H_2$ and $H_3$.

Notice that in the approaches described above we limited our attention to the the problem of linearly separating two sets of points; nonetheless, those formulations can be also applied to the multiclass separation problem (with number of classes $\kappa > 2$) by iteratively solving a sequence of two classes separation problems. Examples of these heuristic methods are the *one-versus-all* and *one-versus-one* schemes (see [2]). While the former approach detects $\kappa - 1$ classifiers, each of which solves the problem of separating points in a particular class from all the points not in that class, the latter alternative computes $\kappa(\kappa - 1)/2$ classifiers, one for every possible pair of classes.

## 5. Numerical Results

In this section, we evaluate the performance of robust and distributionally robust optimization models compared to their deterministic counterparts. The proposed SVM formulations are tested on ten real-world databases, all of which are publicly available and can be downloaded from [29]. The data sets used are listed in Table 2, where the number of features $n \in [4, 279]$, while the number of observations considered $I + J \in [68, 4{,}435]$. For multiclass data sets, we adopted the *one-versus-all* scheme and detected the classifier separating the first class from the remaining ones. This was done to ensure a fair comparison with the results reported in [10], where the same approach was implemented. Clearly, models presented in Section 4 could be also used to identify the remaining $\kappa - 2$ hyperplanes under the *one-versus-all* scheme as well as the $\kappa(\kappa - 1)/2$ classifiers of the *one-versus-one* technique.

The computations have been performed on a 64-bit machine with 8 GB of RAM, a 1.8 GHz Intel i7 processor, and numerical results are obtained under MATLAB environment using MOSEK solver (version 8.1.0.72).

| *Data set* | *Application Field* | *Observations* | *Features* | *Class Balancing* |
|---|---|---|---|---|
| Arrhythmia | Life Sciences | 68 | 279 | 70.59% – 29.41% |
| Breast Cancer | Life Sciences | 683 | 9 | 65.89% – 34.11% |
| Breast Cancer Diagnostic | Life Sciences | 569 | 30 | 62.74% – 37.26% |
| Dermatology | Life Sciences | 358 | 34 | 68.99% – 31.01% |
| Heart Disease | Life Sciences | 297 | 13 | 53.87% – 46.13% |
| Parkinson | Life Sciences | 195 | 22 | 75.38% – 24.62% |
| Climate Model Crashes | Physical Sciences | 540 | 18 | 91.48% – 8.52% |
| Landsat Satellite | Physical Sciences | 4,435 | 36 | 95.47% – 4.53% |
| Ozone Level Detection One | Physical Sciences | 1,848 | 72 | 96.92% – 3.08% |
| Blood Transfusion | Business | 748 | 4 | 76.20% – 23.80% |

Table 2: Summary of data sets from UCI Machine Learning Repository.

For every data set, we first split the overall number of observations $(I+J)$ at our disposal into two disjoint subsets: the former (called *training set*) contains 75% of the observations (of which $I_{\text{tr}}$ belong to the first class and $J_{\text{tr}}$ to the second), the latter (called *testing set*) contains what is left ($I_{\text{ts}} + J_{\text{ts}}$ observations). The observations of the training set are randomly chosen with the only requirement of maintaining the original class balancing, a partition strategy known in the literature as proportional (or stratified) random sampling, *i.e.*:

$$\frac{I_{\text{tr}}}{I_{\text{tr}}+J_{\text{tr}}} = \frac{I}{I+J} \quad \text{and} \quad \frac{J_{\text{tr}}}{I_{\text{tr}}+J_{\text{tr}}} = \frac{J}{I+J}.$$

We refer the reader to [23] for a deeper discussion on proportional random sampling steady performances. For the sake of illustration, we show how to construct training sets on the data set "Breast Cancer Diagnostic". This database lists in total $I+J=569$ observations, of which $I=212$ represent malignant instances and $J=357$ are observations of benign tumors. The class balancing is therefore $62.74\% - 37.26\%$. In the generation of the training set we randomly select $I_{\text{tr}}+J_{\text{tr}}=427$ observations (75% of 569), with $I_{\text{tr}}=268$ belonging to the malignant group and $J_{\text{tr}}=159$ to the benign one. By doing so the class balancing is not altered. Worth noticing that in our computational experiments we did not implement any feature reduction algorithm (such as *feature selection* or *feature extraction*), which means that if an observation belongs to the training set, the entirety of its features will be considered during the training phase. Nonetheless, including such dimensionality reduction approaches could constitute a promising future research direction. Once the partition procedure is complete, different final separating hyperplanes are obtained solving, sequentially, the deterministic (2)-(4), box robust (8)-(9), ellipsoidal robust (12)-(13) and distributionally robust (29)-(30) formulations over the training set. Specifically, we first set the user-defined penalty parameter $\nu$ equally distributed in log space from $10^{-3}$ to $10^{0}$ with 5 discretization points, similarly to what is done in [59], and $k_{\max} = 10^4$. Then, we solve the deterministic formulation under every candidate value $\nu_i$ with $i \in \{1, \dots, 5\}$, record the hyperplane $H_3^{\nu_i}$ and compute the associated misclassification error $\varepsilon^{\nu_i}$. The final deterministic hyperplane $H_3$ is chosen to be the one minimizing the misclassification error $\varepsilon^{\nu_i}$, *i.e.*, $H_3 = H_3^{\nu^*}$ with $\nu^* \in \arg\min\{\varepsilon^{\nu_1}, \dots, \varepsilon^{\nu_5}\}$. The same procedure is repeated for the box formulation, where we additionally set perturbation vectors $\zeta_{x^{(i)}}$ and $\zeta_{y^{(j)}}$ equal to the standard deviation vectors $\sigma_X$ and $\sigma_Y$ of the training groups $X$ and $Y$, *i.e.*, $\zeta_{x^{(i)}} = \sigma_X,\ i = 1, \dots, I_{\text{tr}}$ and $\zeta_{y^{(j)}} = \sigma_Y,\ j = 1, \dots, J_{\text{tr}}$. Similarly, for the ellipsoidal robust formulation where covariance matrices are given by $\Sigma_{x^{(i)}}^{\frac{1}{2}} = \operatorname{diag}(\sigma_X),\ i = 1, \dots, I_{\text{tr}}$ and $\Sigma_{y^{(j)}}^{\frac{1}{2}} = \operatorname{diag}(\sigma_Y),\ j = 1, \dots, J_{\text{tr}}$. For the distributionally robust model, we first perform PCA on the training sets and fix the parameter $K \in \mathbb{N} \setminus \{0\}$ to tune the maximum deviations allowed along principal directions for each observation. For all our test problems, the results we get with values of $K$ larger than 2 are worsening in terms of accuracy levels. Thus, we use $K \in \{1, 2\}$. Worth recalling that setting $K = 1$ grants more dispersion compared to $K = 2$. For all the robust and distributionally robust formulations, procedures are repeated considering increasing levels of $\rho_X, \rho_Y \in \{0.1, 0.2, 0.3\}$. After detecting the final separating hyperplanes using the training data and under the different formulations, we measure their prediction accuracy by reporting the out-of-sample misclassification error on observations belonging to the testing set (*i.e.*, computing testing errors). In order to get stable results, the experiments are performed over 100 different compositions of the hold-out 75%-25% and results are averaged. Furthermore, the procedure is repeated under different hold-outs: 50%-50& and 25%-75%.

These perturbation assumptions for robust and distributionally robust models imply that all the sources of information about features might follow the same form of uncertainty. This is a simplifying assumption driven by the unavailability of explicit details on input data gathering, especially in the medical field where data often comes from heterogeneous sources (*e.g.*, medical imaging, pathology reports, physician notes, genetic assays, lab results, etc.). Naturally, precise knowledge of special structure of input instances would be desired, as would allow taking into account non-homogeneous sources and would therefore lead to wiser choices of perturbations parameters.

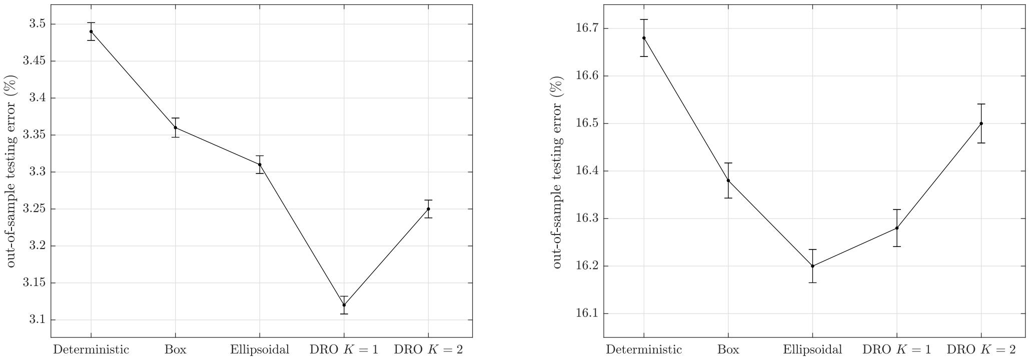
For each formulation and every considered data set, we report in Table 3 mean out-of-sample testing errors and standard deviations[1] for the first hold-out 75%-25%. The solutions under our robust and distributionally robust approaches have intuitive practical appeal, and offer important operational insights. Foremost, by adjusting the radius parameters $\rho_X, \rho_Y$ all robust and distributionally robust formulations are always able to improve prediction accuracies compared to their deterministic counterpart. Therefore, numerical experiments demonstrate that accounting for uncertainty proves to always be beneficial in terms of SVM predictive power.

[1] For each method and every data set, the best result is underlined. Overall, for every single data set, we indicate in bold the lowest out-of-sample testing error rate achieved.

|  | Deterministic | $\rho_X = \rho_Y$ | Box | Ellipsoidal | DRO | DRO |
|---|---|---|---|---|---|---|
|  |  |  | RO | RO | $K = 1$ | $K = 2$ |
| Arrhythmia | 25.65% ± 0.107 | 0.1 | 23.65% ± 0.104 | 24.82% ± 0.102 | 23.65% ± 0.097 | 23.41% ± 0.090 |
|  |  | 0.2 | 23.53% ± 0.092 | 23.06% ± 0.102 | 23.29% ± 0.095 | 23.65% ± 0.093 |
|  |  | 0.3 | 23.06% ± 0.088 | **23.00% ± 0.089** | 23.53% ± 0.090 | 23.65% ± 0.090 |
| Average CPU seconds | 0.560 |  | 0.955 | 1.338 | 125.292 | 104.712 |
| Breast Cancer | 3.49% ± 0.012 | 0.1 | 3.58% ± 0.013 | 3.53% ± 0.012 | 3.34% ± 0.011 | 3.51% ± 0.012 |
|  |  | 0.2 | 3.47% ± 0.012 | 3.43% ± 0.014 | 3.24% ± 0.012 | 3.33% ± 0.011 |
|  |  | 0.3 | 3.36% ± 0.013 | 3.31% ± 0.012 | **3.12% ± 0.012** | 3.25% ± 0.012 |
| Average CPU seconds | 0.244 |  | 0.324 | 1.118 | 7.627 | 7.449 |
| Breast Cancer Diagnostic | 4.89% ± 0.015 | 0.1 | 3.90% ± 0.016 | 4.45% ± 0.015 | 4.66% ± 0.016 | 4.70% ± 0.015 |
|  |  | 0.2 | 3.97% ± 0.015 | **3.89% ± 0.015** | 4.06% ± 0.016 | 4.12% ± 0.017 |
|  |  | 0.3 | 4.04% ± 0.015 | 4.09% ± 0.014 | 4.10% ± 0.015 | 4.23% ± 0.015 |
| Average CPU seconds | 0.261 |  | 0.330 | 0.622 | 20.383 | 17.094 |
| Dermatology | 0.56% ± 0.008 | 0.1 | 0.34% ± 0.007 | 0.34% ± 0.008 | 0.21% ± 0.007 | 0.31% ± 0.008 |
|  |  | 0.2 | 0.24% ± 0.007 | 0.19% ± 0.006 | 0.21% ± 0.007 | 0.30% ± 0.008 |
|  |  | 0.3 | 0.20% ± 0.006 | **0.13% ± 0.005** | 0.29% ± 0.008 | 0.35% ± 0.009 |
| Average CPU seconds | 0.357 |  | 0.608 | 1.072 | 9.958 | 9.331 |
| Heart Disease | 16.68% ± 0.039 | 0.1 | 16.38% ± 0.037 | 16.38% ± 0.036 | 16.28% ± 0.039 | 16.50% ± 0.041 |
|  |  | 0.2 | 17.81% ± 0.045 | **16.20% ± 0.035** | 16.61% ± 0.039 | 16.88% ± 0.037 |
|  |  | 0.3 | 21.57% ± 0.043 | 16.49% ± 0.037 | 18.16% ± 0.040 | 17.32% ± 0.040 |
| Average CPU seconds | 0.228 |  | 0.269 | 1.002 | 3.319 | 3.238 |
| Parkinson | 14.13% ± 0.043 | 0.1 | 13.38% ± 0.032 | **13.00% ± 0.037** | 14.31% ± 0.039 | 14.29% ± 0.039 |
|  |  | 0.2 | 14.42% ± 0.031 | 13.21% ± 0.033 | 13.75% ± 0.038 | 14.00% ± 0.038 |
|  |  | 0.3 | 15.50% ± 0.037 | 13.79% ± 0.033 | 13.60% ± 0.035 | 13.94% ± 0.036 |
| Average CPU seconds | 0.212 |  | 0.314 | 0.611 | 2.851 | 2.811 |
| Climate Model Crashes | 4.99% ± 0.016 | 0.1 | 4.80% ± 0.013 | 4.67% ± 0.013 | **4.34% ± 0.017** | 4.41% ± 0.013 |
|  |  | 0.2 | 6.01% ± 0.011 | 4.48% ± 0.013 | 4.38% ± 0.016 | 4.76% ± 0.019 |
|  |  | 0.3 | 8.50% ± 0.004 | 4.61% ± 0.014 | 5.18% ± 0.015 | 5.62% ± 0.021 |
| Average CPU seconds | 0.252 |  | 0.317 | 0.540 | 8.234 | 8.002 |
| Landsat Satellite | 0.43% ± 0.001 | 0.1 | 0.44% ± 0.002 | 0.42% ± 0.001 | 0.46% ± 0.002 | **0.36% ± 0.001** |
|  |  | 0.2 | 0.42% ± 0.002 | 0.39% ± 0.001 | 0.37% ± 0.001 | 0.40% ± 0.001 |
|  |  | 0.3 | 0.43% ± 0.002 | 0.41% ± 0.002 | 0.37% ± 0.001 | 0.49% ± 0.001 |
| Average CPU seconds | 0.906 |  | 1.041 | 1.250 | 1,142.028 | 1,128.582 |
| Ozone Level Detection One | 6.19% ± 0.013 | 0.1 | 5.32% ± 0.012 | 4.97% ± 0.009 | 4.80% ± 0.012 | 3.15% ± 0.001 |
|  |  | 0.2 | 4.84% ± 0.008 | 3.86% ± 0.007 | 4.11% ± 0.007 | 3.06% ± 0.001 |
|  |  | 0.3 | 4.57% ± 0.008 | 3.81% ± 0.004 | 3.72% ± 0.006 | **3.06% ± 0.001** |
| Average CPU seconds | 0.628 |  | 0.819 | 0.993 | 683.121 | 677.719 |
| Blood Transfusion | 23.49% ± 0.026 | 0.1 | 23.21% ± 0.010 | 23.28% ± 0.013 | 22.87% ± 0.013 | 23.02% ± 0.015 |
|  |  | 0.2 | 23.43% ± 0.007 | **22.55% ± 0.010** | 22.78% ± 0.014 | 22.80% ± 0.014 |
|  |  | 0.3 | 23.53% ± 0.008 | 23.36% ± 0.005 | 23.46% ± 0.021 | 23.09% ± 0.016 |
| Average CPU seconds | 0.255 |  | 0.305 | 0.927 | 7.158 | 7.040 |

Table 3: Average out-of-sample testing errors and standard deviations over 100 runs of the deterministic, robust and distributionally robust models, for the different considered data sets. Hold-out 75%-25%.

Furthermore, it can be noted that once the optimal degree of conservatism is identified, then departing from it translates into a progressive worsening of performances. For instance, for the data set "Breast Cancer" the highest-accuracy model is the distributionally robust with $K = 1$ (out-of-sample testing error rate equal to 3.12%). It follows that opting for progressively more conservative models (ellipsoidal and box robust, in the order) gradually increases out-of-sample testing errors (3.31% and 3.36%, respectively); same conclusion can be drawn solving a less conservative model (distributionally robust with $K = 2$, with an out-of-sample testing error rate setting around 3.25%). For ease of visualization, Figure 5a reports the lowest out-of-sample testing error rate achieved by every formulation under the "Breast Cancer" data set (data from Table 3). In a similar fashion, for the data set "Heart Disease" the highest-accuracy model happens to be the ellipsoidal robust (out-of-sample testing error rate equal to 16.20%) and solving less conservative models (distributionally robust with $K = 1, 2$) gradually increases out-of-sample testing errors (16.28% and 16.50%, respectively); same conclusion can be drawn solving a more conservative model (box robust, with an out-of-sample testing error rate setting around 16.38%), see Figure 5b. The same trends are confirmed under every data set, whose plots are reported in Appendix B (see Figure 8) for the sake of exposition.



(a) Breast Cancer

(b) Heart Disease

Figure 5: Lowest out-of-sample testing error rates over changes of $\rho_X, \rho_Y$ per formulation under the data sets: (a) Breast Cancer; (b) Heart Disease. Vertical error bars represents standard errors. Data of Table 3.

Furthermore, we compare the performance of our models with the accuracy scores reported in [10], that we consider literature benchmark results for robust classification with feature uncertainty. Such comparison highlights that our classifiers perform favorably relative to the standard SVM feature-robust formulation for the majority of the considered problems: 8 out of 10 data sets, as shown in Table 4. Overall, experimental results show that the robustification of the deterministic formulation (2)-(4) proposed in [59] leads to more powerful decision boundaries compared to classical approaches.

| *Data set* | *Table 3* | *Ref. [10]* |
|---|---|---|
| Arrhythmia | **23.00%** | 29.23% |
| Breast Cancer | **3.12%** | 4.26% |
| Breast Cancer Diagnostic | **3.89%** | 4.04% |
| Dermatology | **0.13%** | 1.13% |
| Heart Disease | **16.20%** | 16.61% |
| Parkinson | **13.00%** | 16.41% |
| Climate Model Crashes | 4.34% | **4.07%** |
| Landsat Satellite | **0.36%** | 1.87% |
| Ozone Level Detection One | 3.06% | **2.98%** |
| Blood Transfusion | **22.55%** | 23.62% |

Table 4: Out-of-sample testing error rates comparison. Data of Table 3 against accuracy scores from [10]. For each data set, we indicate in bold the lowest out-of-sample testing error rate achieved.

In Tables 7 and 8 (see Appendix B) we present the results under the 50%-50% and 25%-75% hold-outs. We observe that, with respect to the 75%-25% hold-out, robust and distributionally robust methods significantly outperform the deterministic formulation in terms of prediction accuracy with improvements that increase as the training sample size decreases. This confirms that robust and distributionally robust methods produce high-quality classifiers when the uncertainty increases during the training phase, and therefore their ability to recover the truth from the data increases. To this end, Table 5 shows the robust and distributionally robust improvements in out-of-sample testing errors over their deterministic counterpart. For every data set, we report the best performing model under each hold-out with its average out-of-sample testing error, which we refer to as $\tau^*$. We also compute the improvement ratios $\delta$ as follows:

$$\delta := \frac{\tau^{\text{det}} - \tau^*}{\tau^{\text{det}}}$$

with $\tau^{\text{det}}$ being the average out-of-sample testing error of the deterministic model of each data set.

| | 75%-25% | | | 50%-50% | | | 25%-75% | | |
|---|---|---|---|---|---|---|---|---|---|
| | BEST MODEL | $\tau^*$ | $\delta$ | BEST MODEL | $\tau^*$ | $\delta$ | BEST MODEL | $\tau^*$ | $\delta$ |
| Arrhythmia | Ellipsoidal | 23.00% | 10.32% | Box | 24.00% | 10.33% | Box | 29.04% | 12.47% |
| Breast Cancer | DRO $K = 1$ | 3.12% | 10.54% | DRO $K = 1$ | 3.28% | 11.30% | Box | 3.74% | 22.25% |
| Breast Cancer Diagnostic | Ellipsoidal | 3.89% | 20.45% | Ellipsoidal | 4.41% | 21.25% | Ellipsoidal | 4.94% | 22.20% |
| Dermatology | Ellipsoidal | 0.13% | 76.79% | Ellipsoidal | 0.21% | 76.67% | Ellipsoidal | 0.46% | 77.34% |
| Heart Disease | Ellipsoidal | 16.20% | 2.88% | Ellipsoidal | 17.82% | 4.91% | Ellipsoidal | 19.67% | 5.43% |
| Parkinson | Ellipsoidal | 13.00% | 7.96% | Ellipsoidal | 14.23% | 8.91% | Ellipsoidal | 16.26% | 9.29% |
| Climate Model Crashes | DRO $K = 1$ | 4.34% | 13.03% | Ellipsoidal | 4.87% | 13.19% | Box | 6.40% | 15.90% |
| Landsat Satellite | DRO $K = 2$ | 0.36% | 17.17% | DRO $K = 2$ | 0.41% | 18.58% | DRO $K = 1$ | 0.47% | 20.88% |
| Ozone Level Detection One | DRO $K = 2$ | 3.06% | 50.52% | DRO $K = 1$ | 3.06% | 51.27% | DRO $K = 1$ | 3.08% | 52.09% |
| Blood Transfusion | Ellipsoidal | 22.55% | 4.00% | Ellipsoidal | 22.88% | 4.39% | Ellipsoidal | 23.03% | 4.40% |

Table 5: Robust improvements with respect to the deterministic model on hold-outs 75%-25%, 50%-50%, 25%-75%.

To highlight the statistical significance of our results, under each data set, we also display the $p$-values for the best performing method against the result of its deterministic counterpart, see Table

6. Reported $p$-values are calculated performing a paired-sample $t$-test under the assumption of the null hypothesis that the difference in accuracy of the deterministic and robust or distributionally robust classifier is zero. All results are found to be significant with respect to the typical 5% threshold, except for the "Heart Disease" with 75%-25% hold-out that starts rejecting the null hypothesis at a significance level equal to 8.73%. We recall that the smaller the $p$-value, the more significant is the difference in accuracy.

| | 75%-25% | | 50%-50% | | 25%-75% | |
|---|---|---|---|---|---|---|
| | BEST MODEL | $p$-value | BEST MODEL | $p$-value | BEST MODEL | $p$-value |
| Arrhythmia | Ellipsoidal | 3.26E-02 | Box | 5.69E-04 | Box | 3.60E-09 |
| Breast Cancer | DRO $K = 1$ | 1.38E-02 | DRO $K = 1$ | 1.60E-05 | Box | 3.49E-11 |
| Breast Cancer Diagnostic | Ellipsoidal | 2.54E-11 | Ellipsoidal | 1.39E-17 | Ellipsoidal | 1.54E-15 |
| Dermatology | Ellipsoidal | 9.50E-06 | Ellipsoidal | 4.66E-14 | Ellipsoidal | 1.20E-17 |
| Heart Disease | Ellipsoidal | 8.73E-02 | Ellipsoidal | 2.84E-04 | Ellipsoidal | 2.26E-05 |
| Parkinson | Ellipsoidal | 2.10E-04 | Ellipsoidal | 5.15E-05 | Ellipsoidal | 1.66E-04 |
| Climate Model Crashes | DRO $K = 1$ | 8.70E-03 | Ellipsoidal | 1.50E-06 | Box | 7.51E-07 |
| Landsat Satellite | DRO $K = 2$ | 6.07E-05 | DRO $K = 2$ | 6.20E-04 | DRO $K = 1$ | 1.38E-05 |
| Ozone Level Detection One | DRO $K = 2$ | 3.30E-43 | DRO $K = 1$ | 3.98E-32 | DRO $K = 1$ | 5.52E-47 |
| Blood Transfusion | Ellipsoidal | 8.02E-05 | Ellipsoidal | 1.70E-09 | Ellipsoidal | 1.18E-11 |

Table 6: $p$-values of the best performing robust model on hold-outs 75%-25%, 50%-50%, 25%-75%.

In Figure 6, for the considered hold-outs, we report the number of data sets for which every formulation gave the lowest out-of-sample testing error rate. Histograms clearly underline that for greater training sets (75% of that overall data) less conservative models tend to perform better with respect to the most conservative model (*i.e.*, box). Conversely, as the cardinality of the training set progressively diminishes (down to 25% of that overall data, under the most extreme circumstance) best predictions are obtained using more conservative models. We recall, indeed, that distributionally robust formulations represent more aggressive approaches, since they extract relevant information on the given data and exploit it to define per group perturbation directions.

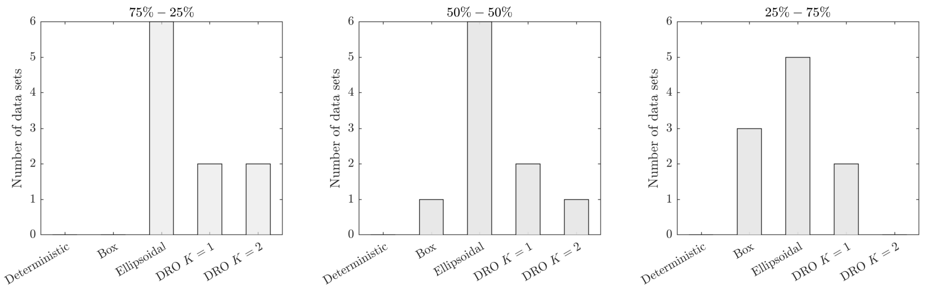


Figure 6: Number of data sets for which every formulation gave the lowest out-of-sample testing error rate. Data of Tables 3, 7, and 8.

To provide advice to final users on when it is valuable to use robust rather than distributionally robust models in practical applications, Figure 7 plots the best performing method against the dimension of the training data set (25%, 50%, 75%). Additionally, the circle sizes are proportional to the values of robust improvements $\delta$ from Table 5. We observe that distributionally robust models outperform robust formulations for the majority of the training sets in the region of high dimensionality (*i.e.*, training sample size with more than approximately 500 observations). Contrariwise, robust models beat distributionally robust methods for the majority of training sets in the region of low dimensionality (*i.e.*, training sample size with less than approximately 500 observations). It is also insightful to compare distributionally robust formulations with distinct degrees of conservatism. Indeed, we observe that more aggressive distributionally robust models (obtained setting $K = 2$) outperform more conservative formulations for the training sets with more than approximately 1,000 observations. This confirms our previous conclusion related to the value of the information at disposal during the training phase, which makes opting for more aggressive models when data might be considered more trustworthy.

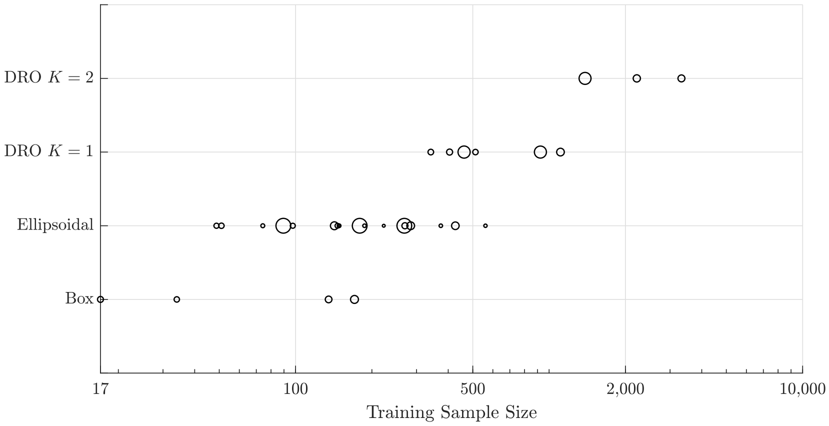


Figure 7: Best performing models versus dimension of the training samples. Data are from Tables 3, 7, and 8. Horizontal axis is in log-scale.

Tables 3, 7, and 8 also present the average CPU time (in seconds) required to find a solution for each method over 500 runs. Numerical results show that solutions for the deterministic and robust formulations were obtained within a few seconds. Contrariwise, higher computational times are observed for the distributionally robust formulations, especially for data sets with larger numbers of observations (*e.g.*, Landsat Satellite and Ozone Level Detection One) or greater number of features (*e.g.*, Arrhythmia). In these cases, deterministic as well as box and ellipsoidal RO methods are several orders of magnitude faster. Therefore, we can conclude that a satisfactory trade-off between accuracy and performing speed is provided by ellipsoidal formulations.

The crucial takeaway message of this work is that hedging against uncertainty in the input observations via robust and distributionally robust approaches offer substantial benefits compared to deterministic formulations and can improve the classification accuracy up to 77.34% (see Table 5). Furthermore, accuracy results recorded by robust and distributionally robust classifiers are more stable, showing less variability when compared to the separators obtained under the deterministic approach. The proposed formulations, overall, allow finding a trade-off between increasing the average performance accuracy and protecting against uncertainty, enabling the decision maker to chose the strategy that is appropriate for each decision making setting.

## 6. Conclusions

In this paper we have presented new optimization models for SVM under uncertainty. Since the consideration of uncertainty is critical to enhance classifiers predictive power, we have formulated robust models with uncertainty regions in the form of both box and ellipsoids, and distributionally robust models that enforce limits on first-order deviations of each input observations along principal directions. We have conducted extensive computational tests on real-world databases with several fields of application. The proposed robust and distributionally robust models have proved to have stronger prediction ability compared to their corresponding deterministic one. Numerical experiments have also shown that as the information at disposal during the training step increases, better prediction accuracy is achieved with more aggressive models (such as the distributionally robust) that account for a higher degree of information. Contrariwise, assuming to have information when such is unreliable has led to poor results. Indeed, as the training sample size gets smaller and the available amount of data is scarce, the utility of implementing distributionally robust approaches has decreased and more conservative models (*i.e.*, box and ellipsoidal robust formulations) have performed better. Overall, taking uncertainty into account during the training phase –to reasonable extents– has always enhanced the classifier's predictive power. Further research activity could be focused on different interesting directions such as: 1) distributionally robust formulations with ellipsoidal supports for SVM; 2) consider the use of different kernel functions for non-linear classifiers under uncertainty; 3) consider the uncertainty in the labels; 4) extend SVM formulations to DRO with different ambiguity sets, such as the ones induced by $\phi$-divergences and Wasserstein distance. On this last regard, the use of distance-based approaches such the Wasserstein-1 metric [71] and an appropriate choice of robustness level could guarantee the inclusion of the true distribution within the ambiguity set with a prescribed level of confidence.

## References

[1] Analui, B., Pflug, G.C., 2014. On distributionally robust multiperiod stochastic optimization. Computational Management Science 11, 197–220.

[2] Anzai, Y., 2012. Pattern recognition and machine learning. Elsevier.

[3] Ardestani-Jaafari, A., Delage, E., 2016. Robust optimization of sums of piecewise linear functions with application to inventory problems. Operations Research 64, 474–494.

[4] Baumann, P., Hochbaum, D.S., Yang, Y.T., 2019. A comparative study of the leading machine learning techniques and two new optimization algorithms. European journal of operational research 272, 1041–1057.

[5] Bayraksan, G., Love, D.K., 2015. Data-driven stochastic programming using phi-divergences, in: The Operations Research Revolution. INFORMS, pp. 1–19.

[6] Ben-Tal, A., Bhadra, S., Bhattacharyya, C., Nath, J.S., 2011. Chance constrained uncertain classification via robust optimization. Mathematical programming 127, 145–173.

[7] Ben-Tal, A., El Ghaoui, L., Nemirovski, A., 2009. Robust optimization. volume 28. Princeton University Press.

[8] Bennett, K.P., Mangasarian, O.L., 1992. Robust linear programming discrimination of two linearly inseparable sets. Optimization methods and software 1, 23–34.

[9] Bertsimas, D., Brown, D.B., Caramanis, C., 2011. Theory and applications of robust optimization. SIAM review 53, 464–501.

[10] Bertsimas, D., Dunn, J., Pawlowski, C., Zhuo, Y.D., 2019. Robust classification. INFORMS Journal on Optimization 1, 2–34.

[11] Bertsimas, D., Gupta, V., Kallus, N., 2018. Data-driven robust optimization. Mathematical Programming 167, 235–292.

[12] Bhadra, S., Nath, J.S., Ben-Tal, A., Bhattacharyya, C., 2009. Interval data classification under partial information: A chance-constraint approach, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer. pp. 208–219.

[13] Bhattacharyya, C., 2004. Robust classification of noisy data using second order cone programming approach. Intelligent Sensing and Information Processing. Proceedings of International Conference , 433–438.

[14] 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, 1073–1089.

[15] Bhattacharyya, C., Pannagadatta, K., Smola, A.J., 2005. A second order cone programming formulation for classifying missing data, in: Neural Information Processing Systems (NIPS), pp. 153–160.

[16] Bi, J., Zhang, T., 2005. Support vector classification with input data uncertainty. Advances in neural information processing systems 17, 161–168.

[17] Biggio, B., Corona, I., Nelson, B., Rubinstein, B.I., Maiorca, D., Fumera, G., Giacinto, G., Roli, F., 2014. Security evaluation of support vector machines in adversarial environments, in: Support Vector Machines Applications. Springer, pp. 105 – 153.

[18] Biggio, B., Nelson, B., Laskov, P., 2011. Support vector machines under adversarial label noise, in: Asian Conference on Machine Learning, pp. 97–112.

[19] Cao, Q., Fu, X., Guo, Y., 2017. Fuzzy chance constrained twin support vector machine for uncertain classification, in: International Conference on Management Science and Engineering Management, Springer. pp. 1508–1521.

[20] Caramanis, C., Mannor, S., 2008. Learning in the limit with adversarial disturbances., in: COLT, Citeseer. pp. 467–478.

[21] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A., 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 408, 189–215.

[22] Ceseracciu, E., Reggiani, M., Sawacha, Z., Sartori, M., Spolaor, F., Cobelli, C., Pagello, E., 2010. SVM classification of locomotion modes using surface electromyography for applications in rehabilitation robotics, in: 19th International Symposium in Robot and Human Interactive Communication, IEEE. pp. 165–170.

[23] Chen, T.Y., Tse, T., Yu, Y.T., 2001. Proportional sampling strategy: A compendium and some insights. Journal of Systems and Software 58, 65–81.

[24] Chernick, M.R., 2011. Bootstrap methods: A guide for practitioners and researchers. volume 619. John Wiley & Sons.

[25] Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

[26] De Cosmis, S., De Leone, R., Kropat, E., Meyer-Nieberg, S., Pickl, S., 2013. Electric load forecasting using support vector machines for robust regression., in: SpringSim (EAIA), p. 9.

[27] Delage, E., Ye, Y., 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations research 58, 595–612.

[28] Dreiseitl, S., Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics 35, 352–359.

[29] Dua, D., Graff, C., 2017. UCI machine learning repository. URL: `http://archive.ics.uci.edu/ml`.

[30] Duchi, J., Namkoong, H., 2019. Variance-based regularization with convex objectives. The Journal of Machine Learning Research 20, 2450–2504.

[31] Duchi, J.C., Namkoong, H., 2021. Learning models with uniform performance via distributionally robust optimization. The Annals of Statistics 49, 1378–1406.

[32] Dudani, S.A., 1976. The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics 4, 325–327.

[33] El Ghaoui, L., Lanckriet, G.R.G., Natsoulis, G., 2003. Robust classification with interval data. Technical Report.

[34] Fan, N., Sadeghi, E., Pardalos, P.M., 2014. Robust support vector machines with polyhedral uncertainty of the input data, in: International Conference on Learning and Intelligent Optimization, Springer. pp. 291–305.

[35] Fujiwara, S., Takeda, A., Kanamori, T., 2017. DC algorithm for extended robust support vector machine. Neural Computation 29, 1406–1438.

[36] Fung, G., Mangasarian, O.L., Shavlik, J.W., 2003. Knowledge-based support vector machine classifiers, in: NIPS, Citeseer. pp. 521–528.

[37] García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: a review. Neural Computing and Applications 19, 263–282.

[38] Ghaoui, L.E., Oks, M., Oustry, F., 2003. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. Operations research 51, 543–556.

[39] Globerson, A., Roweis, S., 2006. Nightmare at test time: Robust learning by feature deletion, ACM. pp. 353–360.

[40] Goh, J., Sim, M., 2010. Distributionally robust optimization and its tractable approximations. Operations research 58, 902–917.

[41] Goldfarb, D., Iyengar, G., 2003. Robust convex quadratically constrained programs. Mathematical Programming 97, 495–515.

[42] Gorissen, B.L., Yanıkoğlu, İ., den Hertog, D., 2015. A practical guide to robust optimization. Omega 53, 124–137.

[43] Gotoh, J., Uryasev, S., 2017. Support vector machines based on convex risk functions and general norms. Annals of Operations Research 249, 301–328.

[44] Gotoh, J.y., Takeda, A., Yamamoto, R., 2014. Interaction between financial risk measures and machine learning methods. Computational Management Science 11, 365–402.

[45] Hashimoto, T., Srivastava, M., Namkoong, H., Liang, P., 2018. Fairness without demographics in repeated loss minimization, in: International Conference on Machine Learning, PMLR. pp. 1929–1938.

[46] Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of educational psychology 24, 417.

[47] Kanamori, T., Fujiwara, S., Takeda, A., 2017. Breakdown point of robust support vector machines. Entropy 19, 83–109.

[48] Katsumata, S., Takeda, A., 2015. Robust cost sensitive support vector machine, in: Artificial intelligence and statistics, PMLR. pp. 434–443.

[49] Khemchandani, R., Chandra, S., et al., 2007. Twin support vector machines for pattern classification. IEEE Transactions on pattern analysis and machine intelligence 29, 905–910.

[50] Kuhn, D., Esfahani, P.M., Nguyen, V.A., Shafieezadeh-Abadeh, S., 2019. Wasserstein distributionally robust optimization: Theory and applications in machine learning, in: Operations Research & Management Science in the Age of Analytics. INFORMS, pp. 130–166.

[51] Kumari, R., Srivastava, S.K., 2017. Machine learning: A review on binary classification. International Journal of Computer Applications 160.

[52] Lanckriet, G.R., Ghaoui, L.E., Bhattacharyya, C., Jordan, M.I., 2002. A robust minimax approach to classification. Journal of Machine Learning Research 3, 555–582.

[53] Le, T., Tran, D., Ma, W., Pham, T., Duong, P., Nguyen, M., 2014. Robust support vector machine, in: 2014 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 4137–4144.

[54] Lee, C., Mehrotra, S., 2015. A distributionally-robust approach for finding support vector machines. Technical Report.

[55] Lee, Y.J., Mangasarian, O.L., 2001. Ssvm: A smooth support vector machine for classification. Computational Optimization and Applications 20, 5–22.

[56] Li, C.N., Shao, Y.H., Deng, N.Y., 2016. Robust l1-norm non-parallel proximal support vector machine. Optimization 65, 169–183.

[57] Li, J., Chen, C., So, A.M.C., 2020. Fast epigraphical projection-based incremental algorithms for wasserstein distributionally robust support vector machine. `arXiv:2010.12865`.

[58] Li, Z., Floudas, C.A., 2012. Robust counterpart optimization: Uncertainty sets, formulations and probabilistic guarantees, in: proceedings of the 6th conference on foundations of computer-aided process operations, Savannah (Georgia).

[59] Liu, X., Potra, F.A., 2009. Pattern separation and prediction via linear and semidefinite programming. Studies in Informatics and Control 18, 71–82.

[60] Liu, Y., Zhang, B., Chen, B., Yang, Y., 2016. Robust solutions to fuzzy one-class support vector machine. Pattern Recognition Letters 71, 73–77.

[61] Livni, R., Crammer, K., Globerson, A., 2012. A simple geometric interpretation of svm using stochastic adversaries, in: Artificial Intelligence and Statistics, PMLR. pp. 722–730.

[62] López, J., Maldonado, S., Carrasco, M., 2017. A robust formulation for twin multiclass support vector machine. Applied Intelligence 47, 1031–1043.

[63] Ma, W., Lejeune, M.A., 2020. A distributionally robust area under curve maximization model. Operations Research Letters 48, 460–466.

[64] Ma, Y., Guo, G., 2014. Support vector machines applications. Springer.

[65] Maldonado, S., López, J., Carrasco, M., 2016. A second-order cone programming formulation for twin support vector machines. Applied Intelligence 45, 265–276.

[66] Marshall, A.W., Olkin, I., 1960. Multivariate Chebyshev inequalities. The Annals of Mathematical Statistics 31, 1001–1014.

[67] Moya, M.M., Hush, D.R., 1996. Network constraints and multi-objective optimization for one-class classification. Neural networks 9, 463–474.

[68] Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A., 2013. Learning with noisy labels, in: Advances in Neural Information Processing Systems, pp. 1196–1204.

[69] Ning, C., You, F., 2017. Data-driven adaptive nested robust optimization: general modeling framework and efficient computational algorithm for decision making under uncertainty. AIChE Journal 63, 3790–3817.

[70] Nocedal, J., Wright, S.J., 2006. Line search methods, in: Numerical optimization. Springer, pp. 30–65.

[71] Noyan, N., Rudolf, G., Lejeune, M., 2018. Distributionally robust optimization with decision-dependent ambiguity set. Technical Report.

[72] Pant, R., Trafalis, T.B., Barker, K., 2011. Support vector machine classification of uncertain and imbalanced data using robust optimization, in: Proceedings of the 15th WSEAS international conference on computers, World Scientific and Engineering Academy and Society (WSEAS) Stevens Point. pp. 369–374.

[73] Pellegrini, M., De Leone, R., Maponi, P., Ferretti, M., 2013. Reducing power consumption in hydrometric level sensor networks using support vector machines., in: Pervasive and Embedded Computing and Communication Systems, pp. 229–232.

[74] Pellegrini, M., De Leone, R., Maponi, P., Rossi, C., 2012. Adaptive sampling for embedded software systems using SVM: Application to water level sensors., in: COMTESSA, editor, Proceedings of the CTW 2012 11th Cologne-Twente Workshop on Graph and Combinatorial Optimization, pp. 100–103.

[75] Pinter, J., 1989. Deterministic approximations of probability inequalities. Zeitschrift für Operations-Research 33, 219–239.

[76] Qi, Z., Tian, Y., Shi, Y., 2013. Robust twin support vector machine for pattern classification. Pattern Recognition 46, 305–316.

[77] Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics 21, 660–674.

[78] Scarf, H., 1958. A min-max solution of an inventory problem, in: Studies in the Mathematical Theory of Inventory and Production. Stanford University Press.

[79] Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L., 2000. New support vector algorithms. Neural computation 12, 1207–1245.

[80] Shaham, U., Yamada, Y., Negahban, S., 2018. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. Neurocomputing 307, 195–204.

[81] Shang, C., Huang, X., You, F., 2017. Data-driven robust optimization based on kernel learning. Computers & Chemical Engineering 106, 464–479.

[82] Shang, C., You, F., 2018. Distributionally robust optimization for planning and scheduling under uncertainty. Computers & Chemical Engineering 110, 53–68.

[83] Shapiro, A., Ahmed, S., 2004. On a class of minimax stochastic programs. SIAM Journal on Optimization 14, 1237–1249.

[84] Shapiro, A., Kleywegt, A., 2002. Minimax analysis of stochastic problems. Optimization Methods and Software 17, 523–542.

[85] Shen, K., Ping, Y., Sun, T., Zhou, Y., 2020. Robust chance constrained optimization with pearson divergence, in: DEStech Transactions on Engineering and Technology Research, pp. 122–125.

[86] 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, 1283–1314.

[87] Silvi, S., Verdenelli, M.C., Cecchini, C., Coman, M.M., Bernabei, M.S., Rosati, J., De Leone, R., Orpianesi, C., Cresci, A., 2014. Probiotic-enriched foods and dietary supplement containing synbio positively affects bowel habits in healthy adults: An assessment using standard statistical analysis and support vector machines. International Journal of Food Sciences & Nutrition 65, 994–1002.

[88] 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, 1359–1385.

[89] Sra, S., Nowozin, S., Wright, S.J., 2012. Optimization for machine learning. Mit Press.

[90] Staib, M., Jegelka, S., 2019. Distributionally robust optimization and generalization in kernel methods. Advances in Neural Information Processing Systems 32, 9134–9144.

[91] Steele, J.M., 2004. The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities. Cambridge University Press.

[92] Stempfel, G., Ralaivola, L., 2009. Learning svms from sloppily labeled data, in: International conference on artificial neural networks, Springer. pp. 884–893.

[93] Takeda, A., Kanamori, T., 2009. A robust approach based on conditional value-at-risk measure to statistical learning problems. European Journal of Operational Research 198, 287–296.

[94] Takeda, A., Sugiyama, M., 2008. $\nu$-support vector machine as conditional value-at-risk minimization, in: Proceedings of the 25th international conference on Machine learning, pp. 1056–1063.

[95] Taskesen, B., Nguyen, V.A., Kuhn, D., Blanchet, J., 2020. A distributionally robust approach to fair classification. arXiv:2007.09530.

[96] Trafalis, T.B., Alwazzi, S.A., 2010. Support vector machine classification with noisy data: a second order cone programming approach. International Journal of General Systems 39, 757–781.

[97] Trafalis, T.B., Gilbert, R.C., 2006. Robust classification and regression using support vector machines. European Journal of Operational Research 173, 893–909.

[98] Trafalis, T.B., Gilbert, R.C., 2007. Robust support vector machines for classification and computational issues. Optimisation Methods and Software 22, 187–198.

[99] Tsyurmasto, P., Gotoh, J., Uryasev, S., 2013. Support vector classification with positive homogeneous risk functionals. Technical Report. Research report 2013-4, Department of Industrial and Systems Engineering.

[100] Utkin, L.V., Chekh, A.I., 2015. A new robust model of one-class classification by interval-valued training data using the triangular kernel. Neural Networks 69, 99–110.

[101] Utkin, L.V., Chekh, A.I., Zhuk, Y.A., 2016. Binary classification svm-based algorithms with interval-valued training data using triangular and epanechnikov kernels. Neural Networks 80, 53–66.

[102] Utkin, L.V., Zhuk, Y.A., 2017. An one-class classification support vector machine model by interval-valued training data. Knowledge-Based Systems 120, 43–56.

[103] Vapnik, V., Chervonenkis, A., 1974. Theory of Pattern Recognition. Nauka, Moscow.

[104] Vishwanathan, S., Murty, M.N., 2002. Ssvm: a simple svm algorithm, in: Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290), IEEE. pp. 2393–2398.

[105] Vitt, C.A., Dentcheva, D., Xiong, H., 2019. Risk-averse classification. Annals of Operations Research, 1–35.

[106] 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.

[107] Wang, X., Fan, N., Pardalos, P.M., 2018. Robust chance-constrained support vector machines with second-order moment information. Annals of Operations Research 263, 45–68.

[108] Wang, Y., 2012. Robust $\nu$-support vector machine based on worst-case conditional value-at-risk minimization. Optimization Methods and Software 27, 1025–1038.

[109] Wang, Y., Nguyen, V.A., Hanasusanto, G.A., 2021. Wasserstein robust support vector machines with fairness constraints. `arXiv:2103.06828`.

[110] Wiesemann, W., Kuhn, D., Sim, M., 2014. Distributionally robust convex optimization. Operations Research 62, 1358–1376.

[111] Wu, Y., Liu, Y., 2007. Robust truncated hinge loss support vector machines. Journal of the American Statistical Association 102, 974–983.

[112] Xiao, H., Biggio, B., Nelson, B., Xiao, H., Eckert, C., Roli, F., 2015. Support vector machines under adversarial label contamination. Neurocomputing 160, 53–62.

[113] Xu, H., Caramanis, C., Mannor, S., 2009a. Robustness and regularization of support vector machines. Journal of Machine Learning Research 10, 1485–1510.

[114] Xu, H., Caramanis, C., Mannor, S., Yun, S., 2009b. Risk sensitive robust support vector machines, in: Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference, IEEE. pp. 4655–4661.

[115] Xu, L., Crammer, K., Schuurmans, D., 2006. Robust support vector machine training via convex outlier ablation, in: Association for the Advancement of Artificial Intelligence, pp. 536–542.

[116] Yue, J., Chen, B., Wang, M.C., 2006. Expected value of distribution information for the newsvendor problem. Operations research 54, 1128–1136.

[117] Žáčková, J., 1966. On minimax solutions of stochastic linear programming problems. Časopis pro Pěstování Matematiky 91, 423–430.

[118] Zhou, L., Wang, L., Liu, L., Ogunbona, P., Shen, D., 2014. Support vector machines for neuroimage analysis: interpretation from discrimination, in: Support Vector Machines Applications. Springer, pp. 191–220.

[119] Zhou, Y., Kantarcioglu, M., Thuraisingham, B., Xi, B., 2012. Adversarial support vector machine learning, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1059–1067.

[120] Zymler, S., Kuhn, D., Rustem, B., 2013. Distributionally robust joint chance constraints with second-order moment information. Mathematical Programming 137, 167–198.

## Appendix A

*Box Robust Formulation*

We now derive the box robust counterpart of formulation (2). Consider the problem:

$$\begin{aligned}
\min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\
\text{s.t.} \quad & a^\top x \le \gamma - 1 + z_{x^{(i)}} && \forall x \in \mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big),\ i = 1,\dots,I \\
& a^\top y \ge \gamma + 1 - z_{y^{(j)}} && \forall y \in \mathcal{U}_{\mathcal{B}}\big(y^{(j)}\big),\ j = 1,\dots,J \\
& z_X \ge 0,\ z_Y \ge 0,
\end{aligned} \tag{31}$$

with:

$$\mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big) := \Big\{ x \in \mathbb{R}^n \ \Big|\ x^{(i)} - \rho_X \zeta_{x^{(i)}} \le x \le x^{(i)} + \rho_X \zeta_{x^{(i)}} \Big\}, \tag{32}$$

$$\mathcal{U}_{\mathcal{B}}\big(y^{(j)}\big) := \Big\{ y \in \mathbb{R}^n \ \Big|\ y^{(j)} - \rho_Y \zeta_{y^{(j)}} \le y \le y^{(j)} + \rho_Y \zeta_{y^{(j)}} \Big\}, \tag{33}$$

where $\zeta_{x^{(i)}} \in \mathbb{R}^n_+$ and $\zeta_{y^{(j)}} \in \mathbb{R}^n_+$ define the perturbation vectors of observations $x^{(i)}$ and $y^{(j)}$, respectively, while $\rho_X \in \mathbb{R}_+$ and $\rho_Y \in \mathbb{R}_+$ are global measures of uncertainty. Formulation (31) can be equivalently re-stated as follows:

$$\begin{aligned}
\min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\
\text{s.t.} \quad & \max_{x \in \mathcal{U}_{\mathcal{B}}(x^{(i)})} \big[a^\top x\big] \le \gamma - 1 + z_{x^{(i)}} && i = 1,\dots,I \\
& \min_{y \in \mathcal{U}_{\mathcal{B}}(y^{(j)})} \big[a^\top y\big] \ge \gamma + 1 - z_{y^{(j)}} && j = 1,\dots,J \\
& z_X \ge 0,\ z_Y \ge 0.
\end{aligned} \tag{34}$$

The left-hand side of the first constraint in (34) can be re-written as follows:

$$\begin{aligned}
\max_{x} \quad & a^\top x \\
\text{s.t.} \quad & x \le x^{(i)} + \rho_X \zeta_{x^{(i)}} \\
& x \ge x^{(i)} - \rho_X \zeta_{x^{(i)}}
\end{aligned}$$

with dual given by:

$$\begin{aligned}
\min_{a^+,\, a^-} \quad & \big(x^{(i)} + \rho_X \zeta_{x^{(i)}}\big)^\top a^+ - \big(x^{(i)} - \rho_X \zeta_{x^{(i)}}\big)^\top a^- \\
\text{s.t.} \quad & a^+ - a^- = a \\
& a^+ \ge 0,\ a^- \ge 0,
\end{aligned}$$

with $a^+$, $a^-$ non-negative dual variables. The dual problem can equivalently be expressed as:

$$\begin{aligned}
\min_{a^+,\, a^-} \quad & (a^+ - a^-)^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top (a^+ + a^-) \\
\text{s.t.} \quad & a^+ - a^- = a \\
& a^+ \geq 0,\ a^- \geq 0,
\end{aligned}$$

which corresponds to:

$$\min_{a}\ a^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top |a|.$$

Therefore the robust linear problem (34) now becomes:

$$\begin{aligned}
\min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\
\text{s.t.} \quad & a^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top |a| \leq \gamma - 1 + z_{x^{(i)}} \quad && i = 1, \ldots, I \\
& \min_{y \in \mathcal{U}_{\mathcal{B}}(y^{(j)})} \big[a^\top y\big] \geq \gamma + 1 - z_{y^{(j)}} && j = 1, \ldots, J \\
& z_X \geq 0,\ z_Y \geq 0.
\end{aligned} \tag{35}$$

Exploiting the following equivalence:

$$\min_{y \in \mathcal{U}_{\mathcal{B}}(y^{(j)})} \big[a^\top y\big] \geq \gamma + 1 - z_{y^{(j)}} \quad \Leftrightarrow \quad \max_{y \in \mathcal{U}_{\mathcal{B}}(y^{(j)})} \big[-a^\top y\big] \leq -\gamma - 1 + z_{y^{(j)}} \quad j = 1, \ldots, J, \tag{36}$$

the same procedure can be followed for the second group of constraints of (34), leading to the following final robust formulation that corresponds to (8):

$$\begin{aligned}
\min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\
\text{s.t.} \quad & a^\top x^{(i)} + \rho_X \zeta_{x^{(i)}}^\top |a| \leq \gamma - 1 + z_{x^{(i)}} \quad && i = 1, \ldots, I \\
& a^\top y^{(j)} - \rho_Y \zeta_{y^{(j)}}^\top |a| \geq \gamma + 1 - z_{y^{(j)}} && j = 1, \ldots, J \\
& z_X \geq 0,\ z_Y \geq 0.
\end{aligned} \tag{37}$$

*Ellipsoidal Robust Formulation*

We now derive the ellipsoidal robust counterpart of formulation (2). Consider the problem:

$$\begin{aligned}
\min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\
\text{s.t.} \quad & a^\top x \leq \gamma - 1 + z_{x^{(i)}} \quad && \forall x \in \mathcal{U}_{\mathcal{E}}\big(x^{(i)}\big),\ i = 1, \ldots, I \\
& a^\top y \geq \gamma + 1 - z_{y^{(j)}} && \forall y \in \mathcal{U}_{\mathcal{E}}\big(y^{(j)}\big),\ j = 1, \ldots, J \\
& z_X \geq 0,\ z_Y \geq 0,
\end{aligned} \tag{38}$$

where:

$$\mathcal{U}_{\mathcal{E}}\big(x^{(i)}\big) := \left\{ x \in \mathbb{R}^n \;\middle|\; \begin{array}{c} x = x^{(i)} + \Sigma_{x^{(i)}}^{\frac{1}{2}} u \\ \|u\|_2 \le \rho_X \end{array} \right\}, \tag{39}$$

$$\mathcal{U}_{\mathcal{E}}\big(y^{(j)}\big) := \left\{ y \in \mathbb{R}^n \;\middle|\; \begin{array}{c} y = y^{(j)} + \Sigma_{y^{(j)}}^{\frac{1}{2}} u \\ \|u\|_2 \le \rho_Y \end{array} \right\}, \tag{40}$$

where $\Sigma_{x^{(i)}} \in \mathbb{R}^{n\times n}$ and $\Sigma_{y^{(j)}} \in \mathbb{R}^{n\times n}$ are positive definite covariance matrices for, respectively, $x^{(i)}$ and $y^{(j)}$, and with the scalars $\rho_X$, $\rho_Y \in \mathbb{R}_+$ denoting the radii of the ellipsoids centered in $x^{(i)}$ and $y^{(j)}$. Equivalently, uncertainty sets (39) and (40) may be expressed as follows:

$$\mathcal{U}_{\mathcal{E}}\big(x^{(i)}\big) := \Big\{ x \in \mathbb{R}^n \;\Big|\; \big(x - x^{(i)}\big)^\top \Sigma_{x^{(i)}}^{-1} \big(x - x^{(i)}\big) \le \rho_X^2 \;\Big\}, \tag{41}$$

$$\mathcal{U}_{\mathcal{E}}\big(y^{(j)}\big) := \Big\{ y \in \mathbb{R}^n \;\Big|\; \big(y - y^{(j)}\big)^\top \Sigma_{y^{(j)}}^{-1} \big(y - y^{(j)}\big) \le \rho_Y^2 \;\Big\}. \tag{42}$$

Once again, we can formulate our problem as:

$$\begin{array}{lll} \min\limits_{a,\gamma,z_X,z_Y} & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) & \\ \text{s.t.} & \max\limits_{x \in \mathcal{U}_{\mathcal{E}}(x^{(i)})} \big[a^\top x\big] \le \gamma - 1 + z_{x^{(i)}} & i = 1, \dots, I \\ & \min\limits_{y \in \mathcal{U}_{\mathcal{E}}(y^{(j)})} \big[a^\top y\big] \ge \gamma + 1 - z_{y^{(j)}} & j = 1, \dots, J \\ & z_X \ge 0,\; z_Y \ge 0. & \end{array} \tag{43}$$

The left-hand side of the first constraint in (43) can be re-written as follows:

$$\begin{array}{ll} \max\limits_{x} & a^\top x \\ \text{s.t.} & x \in \mathcal{U}_{\mathcal{E}}\big(x^{(i)}\big) \end{array}$$

which is equivalent to:

$$\begin{array}{lll} a^\top x^{(i)} + & \max\limits_{u} & \Big[a^\top \Sigma_{x^{(i)}}^{\frac{1}{2}} u\Big] \\ & \text{s.t.} & \|u\|_2 \le \rho_X. \end{array} \tag{44}$$

Applying the Cauchy-Schwarz inequality ([91]) we get: $|a^\top \Sigma_{x^{(i)}}^{\frac{1}{2}} u| \le \|a^\top \Sigma_{x^{(i)}}^{\frac{1}{2}}\|_2 \cdot \|u\|_2$. Therefore, since $\|u\|_2 \le \rho_X$, problem (44) becomes:

$$a^\top x^{(i)} + \|a^\top \Sigma_{x^{(i)}}^{\frac{1}{2}}\|_2 \cdot \|u\|_2 \quad \Leftrightarrow \quad a^\top x^{(i)} + \rho_X \|\Sigma_{x^{(i)}}^{\frac{1}{2}} a\|_2. \tag{45}$$

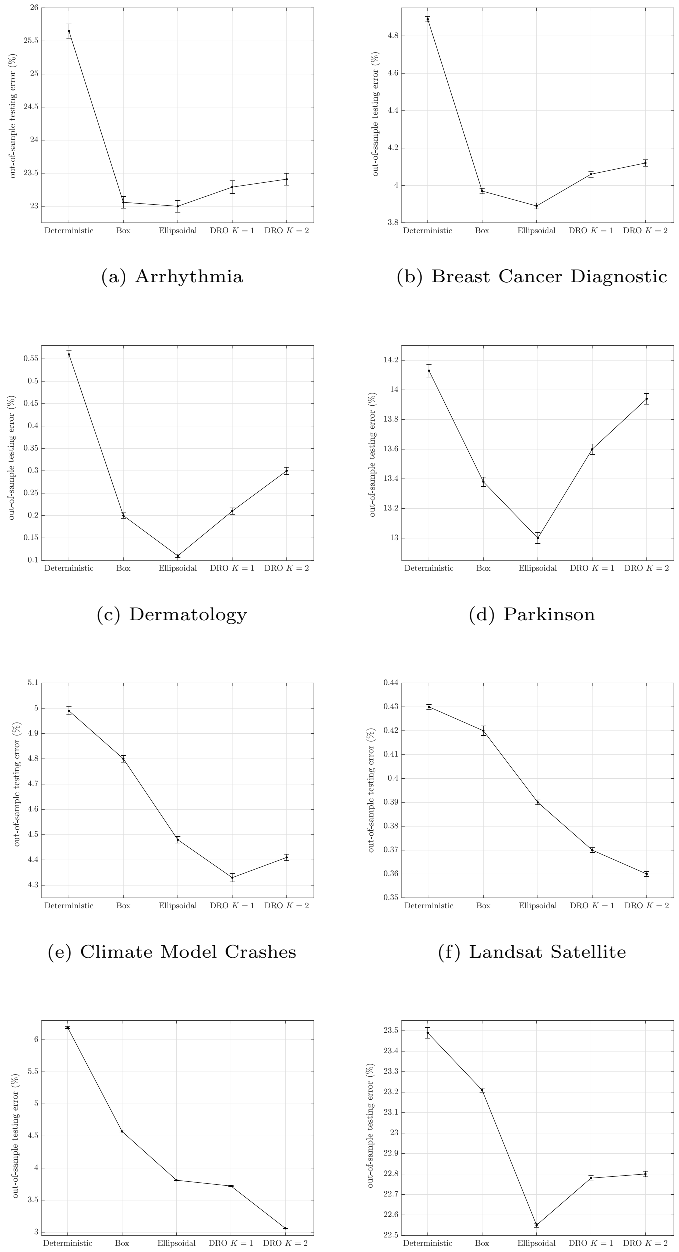
Exploiting the equivalence:

$$\min_{y\in\mathcal{U}_{\mathcal{E}}(y^{(j)})}\big[a^\top y\big] \geq \gamma + 1 - z_{y^{(j)}} \quad\Leftrightarrow\quad \max_{y\in\mathcal{U}_{\mathcal{E}}(y^{(j)})}\big[-a^\top y\big] \leq -\gamma - 1 + z_{y^{(j)}} \quad j = 1,\dots,J, \tag{46}$$

the same procedure can be followed for the second group of constraints of (43), leading to the final ellipsoidal robust formulation given by:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} \quad & a^\top x^{(i)} + \rho_X \|\Sigma_{x^{(i)}}^{\frac{1}{2}} a\|_2 \leq \gamma - 1 + z_{x^{(i)}} \quad i = 1,\dots,I \\ & a^\top y^{(j)} - \rho_Y \|\Sigma_{y^{(j)}}^{\frac{1}{2}} a\|_2 \geq \gamma + 1 - z_{y^{(j)}} \quad j = 1,\dots,J \\ & z_X \geq 0,\ z_Y \geq 0. \end{aligned} \tag{47}$$

# Appendix B



(a) Arrhythmia

(b) Breast Cancer Diagnostic

(c) Dermatology

(d) Parkinson

(e) Climate Model Crashes

(f) Landsat Satellite

(g) Ozone Level Detection One

(h) Blood Transfusion

Figure 8: Lowest out-of-sample testing error rates over changes of $\rho_X, \rho_Y$ per formulation under the data sets: (a) Arrhythmia; (b) Breast Cancer Diagnostic; (c) Dermatology; (d) Parkinson; (e) Climate Model Crashes; (f) Landsat Satellite; (g) Ozone Level Detection One; (h) Blood Transfusion. Vertical error bars represents standard errors. Graphics refer to data of Table 3.

|  | Deterministic | $\rho_X = \rho_Y$ | Box | Ellipsoidal | DRO | DRO |
|---|---|---|---|---|---|---|
|  |  |  | RO | RO | $K = 1$ | $K = 2$ |
| Arrhythmia | 26.76% ± 0.080 | 0.1 | 25.44% ± 0.075 | 25.76% ± 0.075 | 25.44% ± 0.075 | 25.50% ± 0.076 |
|  |  | 0.2 | 25.12% ± 0.072 | 25.00% ± 0.075 | 25.62% ± 0.076 | 25.59% ± 0.075 |
|  |  | 0.3 | **24.00% ± 0.069** | 24.32% ± 0.069 | 25.65% ± 0.074 | 25.59% ± 0.073 |
| Average CPU seconds | 0.433 |  | 0.681 | 0.792 | 36.200 | 32.092 |
| Breast Cancer | 3.70% ± 0.010 | 0.1 | 3.56% ± 0.007 | 3.54% ± 0.008 | 3.42% ± 0.008 | 3.64% ± 0.008 |
|  |  | 0.2 | 3.43% ± 0.007 | 3.40% ± 0.008 | 3.40% ± 0.008 | 3.55% ± 0.008 |
|  |  | 0.3 | 3.32% ± 0.007 | 3.29% ± 0.008 | **3.28% ± 0.007** | 3.38% ± 0.008 |
| Average CPU seconds | 0.240 |  | 0.242 | 0.846 | 5.036 | 4.984 |
| Breast Cancer Diagnostic | 5.60% ± 0.011 | 0.1 | 5.08% ± 0.012 | 4.64% ± 0.010 | 5.18% ± 0.013 | 5.25% ± 0.012 |
|  |  | 0.2 | 5.19% ± 0.014 | 4.45% ± 0.011 | 5.32% ± 0.012 | 5.42% ± 0.012 |
|  |  | 0.3 | 5.29% ± 0.017 | **4.41% ± 0.010** | 5.44% ± 0.012 | 5.39% ± 0.012 |
| Average CPU seconds | 0.260 |  | 0.302 | 0.593 | 9.382 | 9.229 |
| Dermatology | 0.90% ± 0.008 | 0.1 | 0.57% ± 0.008 | 0.81% ± 0.012 | 0.42% ± 0.006 | 0.49% ± 0.008 |
|  |  | 0.2 | 0.41% ± 0.006 | 0.46% ± 0.007 | 0.46% ± 0.007 | 0.47% ± 0.008 |
|  |  | 0.3 | 0.32% ± 0.005 | **0.21% ± 0.004** | 0.47% ± 0.008 | 0.50% ± 0.008 |
| Average CPU seconds | 0.246 |  | 0.255 | 0.617 | 8.198 | 7.283 |
| Heart Disease | 18.74% ± 0.027 | 0.1 | 18.11% ± 0.022 | 18.38% ± 0.027 | 18.23% ± 0.029 | 18.49% ± 0.028 |
|  |  | 0.2 | 19.37% ± 0.033 | 17.82% ± 0.025 | 18.56% ± 0.031 | 18.41% ± 0.029 |
|  |  | 0.3 | 24.57% ± 0.053 | **17.82% ± 0.024** | 19.47% ± 0.032 | 18.68% ± 0.029 |
| Average CPU seconds | 0.227 |  | 0.256 | 0.521 | 2.478 | 2.347 |
| Parkinson | 15.62% ± 0.036 | 0.1 | 14.28% ± 0.031 | 14.55% ± 0.027 | 15.55% ± 0.039 | 15.36% ± 0.037 |
|  |  | 0.2 | 15.57% ± 0.030 | 14.47% ± 0.025 | 15.05% ± 0.035 | 15.29% ± 0.036 |
|  |  | 0.3 | 17.26% ± 0.041 | **14.23% ± 0.023** | 15.22% ± 0.037 | 15.19% ± 0.034 |
| Average CPU seconds | 0.206 |  | 0.244 | 0.504 | 1.845 | 1.788 |
| Climate Model Crashes | 5.61% ± 0.015 | 0.1 | 5.02% ± 0.010 | 5.21% ± 0.012 | 5.34% ± 0.014 | 5.42% ± 0.015 |
|  |  | 0.2 | 6.19% ± 0.011 | 4.90% ± 0.011 | 5.33% ± 0.014 | 5.52% ± 0.014 |
|  |  | 0.3 | 8.46% ± 0.003 | **4.87% ± 0.010** | 5.73% ± 0.014 | 6.37% ± 0.011 |
| Average CPU seconds | 0.239 |  | 0.254 | 0.517 | 5.518 | 5.433 |
| Landsat Satellite | 0.51% ± 0.002 | 0.1 | 0.47% ± 0.001 | 0.44% ± 0.001 | 0.48% ± 0.001 | **0.41% ± 0.001** |
|  |  | 0.2 | 0.47% ± 0.001 | 0.43% ± 0.001 | 0.47% ± 0.002 | 0.43% ± 0.001 |
|  |  | 0.3 | 0.48% ± 0.002 | 0.44% ± 0.001 | 0.42% ± 0.001 | 0.48% ± 0.002 |
| Average CPU seconds | 0.654 |  | 0.684 | 0.846 | 522.252 | 484.864 |
| Ozone Level Detection One | 6.27% ± 0.018 | 0.1 | 5.93% ± 0.015 | 5.81% ± 0.014 | 3.71% ± 0.009 | 5.81% ± 0.015 |
|  |  | 0.2 | 5.79% ± 0.014 | 5.11% ± 0.011 | 3.07% ± 0.001 | 4.47% ± 0.005 |
|  |  | 0.3 | 5.71% ± 0.014 | 4.42% ± 0.008 | **3.06% ± 0.001** | 4.36% ± 0.005 |
| Average CPU seconds | 0.489 |  | 0.502 | 0.667 | 310.425 | 291.384 |
| Blood Transfusion | 23.93% ± 0.016 | 0.1 | 23.15% ± 0.005 | 22.94% ± 0.007 | 23.48% ± 0.011 | 23.53% ± 0.009 |
|  |  | 0.2 | 23.45% ± 0.004 | **22.88% ± 0.006** | 23.55% ± 0.013 | 23.59% ± 0.008 |
|  |  | 0.3 | 23.60% ± 0.003 | 23.44% ± 0.005 | 23.88% ± 0.021 | 23.91% ± 0.016 |
| Average CPU seconds | 0.216 |  | 0.230 | 0.492 | 4.952 | 4.935 |

Table 7: Average out-of-sample testing errors and standard deviations over 100 runs of the deterministic, robust and distributionally robust models, for the different considered data sets. Hold-out 50%-50%. For each method and under every data set, the best result is underlined. Overall, for every single data set, we indicate in bold the lowest out-of-sample testing error rate achieved.

| | Deterministic | $\rho_X = \rho_Y$ | Box | Ellipsoidal | DRO | DRO |
|---|---|---|---|---|---|---|
| | | | RO | RO | $K = 1$ | $K = 2$ |
| Arrhythmia | 33.18% ± 0.068 | 0.1 | 31.12% ± 0.074 | 31.98% ± 0.070 | 32.16% ± 0.083 | 32.06% ± 0.082 |
| | | 0.2 | 29.82% ± 0.069 | 31.37% ± 0.073 | 32.12% ± 0.082 | 32.14% ± 0.082 |
| | | 0.3 | **29.04% ± 0.064** | 30.29% ± 0.075 | 32.06% ± 0.082 | 32.24% ± 0.082 |
| Average CPU seconds | 0.423 | | 0.535 | 0.780 | 7.646 | 6.182 |
| Breast Cancer | 4.81% ± 0.013 | 0.1 | 4.35% ± 0.011 | 4.25% ± 0.009 | 4.47% ± 0.013 | 4.35% ± 0.010 |
| | | 0.2 | 3.96% ± 0.008 | 4.03% ± 0.009 | 4.31% ± 0.011 | 4.22% ± 0.010 |
| | | 0.3 | **3.74% ± 0.007** | 3.81% ± 0.008 | 3.91% ± 0.009 | 3.94% ± 0.009 |
| Average CPU seconds | 0.226 | | 0.240 | 0.608 | 2.621 | 2.552 |
| Breast Cancer Diagnostic | 6.35% ± 0.013 | 0.1 | 5.16% ± 0.009 | 5.17% ± 0.010 | 6.02% ± 0.011 | 6.02% ± 0.011 |
| | | 0.2 | 5.19% ± 0.012 | 4.96% ± 0.009 | 6.00% ± 0.012 | 6.04% ± 0.012 |
| | | 0.3 | 6.03% ± 0.015 | **4.94% ± 0.009** | 6.04% ± 0.011 | 6.06% ± 0.012 |
| Average CPU seconds | 0.250 | | 0.258 | 0.575 | 3.586 | 3.541 |
| Dermatology | 2.03% ± 0.014 | 0.1 | 1.12% ± 0.010 | 1.02% ± 0.011 | 0.66% ± 0.007 | 0.74% ± 0.007 |
| | | 0.2 | 0.76% ± 0.008 | 0.65% ± 0.008 | 0.69% ± 0.008 | 0.74% ± 0.008 |
| | | 0.3 | 0.54% ± 0.006 | **0.46% ± 0.006** | 0.72% ± 0.008 | 0.76% ± 0.007 |
| Average CPU seconds | 0.215 | | 0.239 | 0.590 | 2.148 | 2.130 |
| Heart Disease | 20.90% ± 0.027 | 0.1 | 20.50% ± 0.030 | 20.48% ± 0.026 | 20.05% ± 0.028 | 20.22% ± 0.027 |
| | | 0.2 | 21.14% ± 0.036 | 19.72% ± 0.025 | 20.42% ± 0.029 | 20.60% ± 0.030 |
| | | 0.3 | 23.90% ± 0.045 | **19.67% ± 0.025** | 20.81% ± 0.035 | 20.75% ± 0.032 |
| Average CPU seconds | 0.222 | | 0.229 | 0.492 | 1.414 | 1.269 |
| Parkinson | 17.92% ± 0.044 | 0.1 | 16.67% ± 0.036 | 16.55% ± 0.039 | 17.87% ± 0.041 | 18.12% ± 0.046 |
| | | 0.2 | 17.87% ± 0.039 | 16.39% ± 0.039 | 18.29% ± 0.043 | 17.92% ± 0.042 |
| | | 0.3 | 19.82% ± 0.043 | **16.26% ± 0.035** | 18.47% ± 0.046 | 17.87% ± 0.041 |
| Average CPU seconds | 0.206 | | 0.227 | 0.461 | 1.099 | 1.037 |
| Climate Model Crashes | 7.61% ± 0.022 | 0.1 | **6.40% ± 0.016** | 7.26% ± 0.020 | 7.50% ± 0.024 | 7.88% ± 0.023 |
| | | 0.2 | 6.61% ± 0.012 | 6.90% ± 0.018 | 7.84% ± 0.028 | 9.04% ± 0.032 |
| | | 0.3 | 8.06% ± 0.009 | 6.55% ± 0.015 | 8.51% ± 0.032 | 9.04% ± 0.036 |
| Average CPU seconds | 0.226 | | 0.239 | 0.491 | 2.352 | 2.304 |
| Landsat Satellite | 0.59% ± 0.002 | 0.1 | 0.51% ± 0.001 | 0.50% ± 0.002 | 0.51% ± 0.002 | 0.51% ± 0.002 |
| | | 0.2 | 0.49% ± 0.001 | 0.49% ± 0.002 | 0.50% ± 0.002 | 0.50% ± 0.002 |
| | | 0.3 | 0.53% ± 0.002 | 0.48% ± 0.002 | **0.47% ± 0.002** | 0.49% ± 0.002 |
| Average CPU seconds | 0.414 | | 0.423 | 0.701 | 122.338 | 118.907 |
| Ozone Level Detection One | 6.43% ± 0.012 | 0.1 | 6.26% ± 0.017 | 6.20% ± 0.018 | 4.25% ± 0.010 | 5.12% ± 0.017 |
| | | 0.2 | 5.08% ± 0.014 | 5.33% ± 0.017 | 3.32% ± 0.006 | 4.74% ± 0.017 |
| | | 0.3 | 5.06% ± 0.011 | 4.70% ± 0.010 | **3.08% ± 0.001** | 4.73% ± 0.019 |
| Average CPU seconds | 0.393 | | 0.404 | 0.612 | 97.110 | 96.562 |
| Blood Transfusion | 24.09% ± 0.014 | 0.1 | 23.43% ± 0.006 | 23.17% ± 0.005 | 23.42% ± 0.028 | 23.61% ± 0.011 |
| | | 0.2 | 23.57% ± 0.004 | **23.03% ± 0.008** | 23.25% ± 0.036 | 24.13% ± 0.021 |
| | | 0.3 | 23.66% ± 0.002 | 23.45% ± 0.005 | 23.43% ± 0.045 | 25.42% ± 0.026 |
| Average CPU seconds | 0.214 | | 0.223 | 0.490 | 2.622 | 2.587 |

Table 8: Average out-of-sample testing errors and standard deviations over 100 runs of the deterministic, robust and distributionally robust models, for the different considered data sets. Hold-out 25%-75%. For each method and under every data set, the best result is underlined. Overall, for every single data set, we indicate in bold the lowest out-of-sample testing error rate achieved.