Ann Oper Res
DOI 10.1007/s10479-015-2039-6

Robust chance-constrained support vector machines with second-order moment information

Ximing Wang[1] · **Neng Fan**[2] · **Panos M. Pardalos**[1]

Abstract Support vector machines (SVM) is one of the well known supervised classes of learning algorithms. Basic SVM models are dealing with the situation where the exact values of the data points are known. This paper studies SVM when the data points are uncertain. With some properties known for the distributions, chance-constrained SVM is used to ensure the small probability of misclassification for the uncertain data. As infinite number of distributions could have the known properties, the robust chance-constrained SVM requires efficient transformations of the chance constraints to make the problem solvable. In this paper, robust chance-constrained SVM with second-order moment information is studied and we obtain equivalent semidefinite programming and second order cone programming reformulations. The geometric interpretation is presented and numerical experiments are conducted. Three types of estimation errors for mean and covariance information are studied in this paper and the corresponding formulations and techniques to handle these types of errors are presented.

Keywords Support vector machines · Robust chance constraints · Semidefinite programming · Second order cone programming · Second-order moment information · Estimation errors

✉ Panos M. Pardalos
pardalos@ufl.edu

Ximing Wang
x.wang@ufl.edu

Neng Fan
nfan@email.arizona.edu

[1] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA

[2] Department of Systems and Industrial Engineering, University of Arizona, Tucson, AZ 85721, USA

Published online: 12 October 2015

1 Introduction

In recent years, machine learning and data mining have an explosive growth with new developments in science and technology. Many techniques have been proposed to deal with different datasets. The essentials of most techniques are optimization problems. Traditional machine learning models are dealing with data when the exact values are known. This paper considers the case when uncertainties exist in data.

As one of the well known supervised learning algorithms, Support Vector Machines (SVM) is gaining more and more attention. It was proposed by Vapnik (1998, 1999) as a maximum-margin classifier. Tutorials on SVM could be found in Burges (1998), Abe (2010), Ben-Hur and Weston (2010), Chang and Lin (2011). In recent years, SVM has been applied to many fields and has many algorithmic and modeling variations (Tian et al. 2012; Wang and Pardalos 2014).

Basic SVM models are dealing with the situation where the exact values of the data points are known. When the data points are uncertain, different models have been proposed to formulate the SVM with uncertainties. Bi and Zhang (2005) assumed the data points are subject to an additive noise which is bounded by norm and proposed a very direct model. However, this model cannot guarantee a generally good performance on the uncertainty set. To guarantee an optimal performance when the worst-case scenario constraints are still satisfied, robust optimization is utilized. Trafalis and Gilbert (2006, 2007), Trafalis and Alwazzi (2010), Pant et al. (2011), Xanthopoulos et al. (2012) proposed a robust optimization model when the perturbation of the uncertain data is bounded by norm, where some efficient linear programming models are presented under certain conditions. Xanthopoulos et al. (2014) studied robust generalized eigenvalue classifier with ellipsoidal uncertainty. Ghaoui et al. (2003) derived a robust model when the uncertainty is expressed as intervals with support and extremum values. Fan et al. (2014) studied a more general case for polyhedral uncertainties.

Robust optimization is also used for solving SVM with chance constraints to ensure the small probability of misclassification for the uncertain data. The chance constraints are transformed by different bounding inequalities, for example multivariate Chebyshev inequality (Bhattacharyya et al. 2004; Shivaswamy et al. 2006) and Bernstein bounding schemes (Ben-Tal et al. 2011). The Chebyshev based model employs moment information of the uncertain training points. The Bernstein bounds can be less conservative than the Chebyshev bounds since it employs both support and moment information, but it also makes a strong assumption that all the elements in the data set are independent.

This paper studies the reformulation of robust chance-constrained SVM into equivalent Semidefinite Programming (SDP) model and Second Order Cone Programming (SOCP) model with second-order moment information of the uncertain data provided. Since the moment information is often unknown but to be estimated from the data, there might be estimation errors. This paper also considers different estimation errors and proposes corresponding models. Comparing with literatures that have studied the chance-constrained SVM, this paper proposes a different proof for obtaining the equivalent SOCP formulation. Similar result has been obtained based on the Chebyshev inequality (Bhattacharyya et al. 2004; Shivaswamy et al. 2006), while our paper presents an approach from optimization prospective. A new and equivalent formulation based on SDP is proposed in this paper for the robust chance-constrained SVM. This SDP builds a bridge between robust chance-constrained SVM and SOCP. Besides these, three types of errors for mean and covariance information are studied in this paper. Comparing with Bhattacharyya et al. (2004), Shivaswamy et al. (2006) for fixed mean and covariance and Ben-Tal et al. (2011) for mean and the expectation of the square

of each element (not general covariance matrix since it does not consider the covariance between different elements in the feature vector), this paper considers more general cases, and presents the corresponding formulations and techniques to handle these three types of errors.

The structure of this paper is as follows: Sect. 2 consists of an introduction of the robust chance-constrained SVM model. Section 3 describes the reformulation of robust chance-constrained SVM into equivalent SDP and SOCP models, as well as the geometric interpretation. Section 4 discusses the estimation errors and corresponding models, performance measures are also discussed in this section. Section 5 contains the numerical experiments on the equivalence and on the estimation and performance issues. Section 6 concludes this paper.

2 Robust chance-constrained SVM

SVM constructs maximum-margin classifiers, such that small perturbations in data are least likely to cause misclassification. Empirically, SVM works really well and is a well-known supervised learning algorithm proposed by Vapnik (1998, 1999). Suppose a two-class dataset of m data points $\{\mathbf{x}_i, y_i\}_{i=1}^m$ with n-dimensional features $\mathbf{x}_i \in \mathbb{R}^n$ and respective class labels $y_i \in \{+1, -1\}$. For linearly separable datasets, there exists a hyperplane $\mathbf{w}^\top \mathbf{x} + b = 0$ to separate the two classes and the corresponding classification rule is based on the sign$(\mathbf{w}^\top \mathbf{x} + b)$. If this value is positive, $\mathbf{x}$ is classified to be in $+1$ class; otherwise, -1 class.

The datapoints that the margin pushes up against are called support vectors. A maximum-margin hyperplane is one that maximizes the distance between the hyperplane and the support vectors. For the separating hyperplane $\mathbf{w}^\top \mathbf{x} + b = 0$, $\mathbf{w}$ and b could be normalized so that $\mathbf{w}^\top \mathbf{x} + b = +1$ goes through the support vectors of $+1$ class, and $\mathbf{w}^\top \mathbf{x} + b = -1$ goes through the support vectors of -1 class. The distance between these two hyperplane, i.e., the margin width, is $\frac{2}{\|\mathbf{w}\|_2^2}$, therefore, maximization of the margin can be performed as minimization of $\frac{1}{2}\|\mathbf{w}\|_2^2$ subject to separation constraints. This can be expressed as the following quadratic optimization problem:

$$\textbf{(SVM)}$$

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 \tag{1a}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i = 1, \ldots, m \tag{1b}$$

The above is valid in the case that the two classes are linearly separable. When they are not, mislabeled samples need to be allowed where soft margin SVM arises. Soft margin SVM introduces non-negative slack variables ξ_i to measure the distance of within-margine or misclassified data $\mathbf{x}_i$ to the hyperplane with the correct label, and $\xi_i = \max\{0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\}$. When $0 < \xi_i < 1$, the data is within margine but correctly classified; when $\xi_i > 1$, the data is misclassified. The objective function is then adding a term that penalizes these slack variables, and the optimization is a trade off between a large margin and a small error penalty. The soft margin SVM formulation with L_1 regularization (Cortes and Vapnik 1995) is:

$$\textbf{(SVM} - \textbf{SoftMargin)}$$

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^m \xi_i \tag{2a}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, m \tag{2b}$$

where C is a trade-off parameter.

The above soft margin SVM can be reformulated into second order cone program by replacing the term $\|\mathbf{w}\|_2^2$ in the objective function by a constraint upper bounding $\|\mathbf{w}\|_2$ by a constant W (Shivaswamy et al. 2006). This would imply:

$$\min_{\mathbf{w},b,\xi_i} \sum_{i=1}^{m} \xi_i \tag{3a}$$

$$\text{s.t. } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, m \tag{3b}$$

$$||\mathbf{w}||_2 \leq W \tag{3c}$$

The constraint $||\mathbf{w}||_2 \leq W$ is a second order cone constraint, the objective function and other constraints are linear. So this formulation is SOCP. The difference with the original formulation is that a direct bound W is put on $\|\mathbf{w}\|_2$ instead of the trade-off parameter C on slack variables in the objective function. It can be shown that when choosing C and W properly, model (3) and model (2) give the same optimal values of $(\mathbf{w}, b, \xi_i)$. Therefore, in the following, **(SVM-SoftMargin)** is taken to be the basic model which could be transformed into SOCP implicitly.

When uncertainties exist in the data points, the model needs to be modified to contain the uncertainty information. Suppose there are m training data points in $\mathbb{R}^n$, use $\tilde{\mathbf{x}}_i = [\tilde{x}_{i1}, \ldots, \tilde{x}_{in}]^\top \in \mathbb{R}^n, i = 1, \ldots, m$ to denote the uncertain training data points and $y_i \in \{+1, -1\}, i = 1, \ldots, m$ to denote the respective class labels. The soft margin SVM formulation is as following:

$$(\mathbf{SVM - Uncertainty})$$

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m} \xi_i \tag{4a}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, m \tag{4b}$$

The Chance-Constrained Program (CCP) is to ensure the small probability of misclassification for the uncertain data. The chance-constrained SVM formulation is:

$$(\mathbf{SVM - CCP})$$

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m} \xi_i \tag{5a}$$

$$\text{s.t.} \quad \mathbb{P}\Big\{y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \leq 1 - \xi_i\Big\} \leq \varepsilon, \quad \xi_i \geq 0, \quad i = 1, \ldots, m \tag{5b}$$

where $0 < \varepsilon < 1$ is a prameter given by the user and close to 0, $\mathbb{P}\{\cdot\}$ is the probability distribution. This model ensures an upper bound on the misclassification probability, but the chance constraints are typically non-convex so the problem is very hard to solve.

In practice, the exact probability distribution of the random variables are often unknown and hard to get. Only some properties of the distribution could be acquired, such as the first and second moments. To deal with the uncertainty in probability distribution, the distributionally robust or ambiguous chance constraint is developed and adopted to represent a conservative approximation of the original problem. Let $\mathcal{P}$ be the set of all probability distributions that have the known properties of $\mathbb{P}$, then the distributionally robust chance-constrained SVM formulation is Shivaswamy et al. (2006), Ben-Tal et al. (2011):

$$\textbf{(SVM-RCCP)}$$

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \tag{6a}$$

$$\text{s.t.} \quad \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\Big\{y_i(\mathbf{w}^\top\tilde{\mathbf{x}}_i + b) \le 1-\xi_i\Big\} \le \varepsilon, \quad \xi_i \ge 0, \quad i=1,\ldots,m \tag{6b}$$

It is easy to see that if the distributionally robust chance constraint in **(SVM-RCCP)** is satisfied, then the chance constraint in **(SVM-CCP)** will also be satisfied under the true probability distribution.

3 Reformulation of (SVM-RCCP) into SDP and SOCP

3.1 Reformulation of (SVM-RCCP) into SDP

Assume the first and second moment information of the random variables $\tilde{\mathbf{x}}_i$ are known. Let $\boldsymbol{\mu}_i = \mathbf{E}[\tilde{\mathbf{x}}_i] \in \mathbb{R}^n$ be the mean vector and $\boldsymbol{\Sigma}_i = \mathbf{E}\big[(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_i)(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_i)^\top\big] \in \mathbb{S}^n$ be the covariance matrix of the random variable $\tilde{\mathbf{x}}_i$. Let $\mathcal{P}$ be the set of all probability distributions that have the same first and second moments. We have the following theorem.

Theorem 1 *The robust chance-constrained SVM model* **(SVM-RCCP)** *can be reformulated as the following SDP model:*

$$\textbf{(SVM-SDP)}$$

$$\min_{\mathbf{w},b,\xi_i,\mathbf{N}_i,\alpha_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \tag{7a}$$

$$s.t. \quad \alpha_i - \frac{1}{\varepsilon}Trace(\Omega_i\mathbf{N}_i) \ge 0, \quad \xi_i \ge 0 \tag{7b}$$

$$\mathbf{N}_i \succeq 0, \quad \mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_ib + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0 \tag{7c}$$

where $\Omega_i = \begin{bmatrix} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^\top & 1 \end{bmatrix}$*; and for matrix* $\mathbf{A}$, $\mathbf{A} \succeq 0$ *means* $\mathbf{A}$ *is positive semidefinite.*

Proof Similar to Zymler et al. (2013), for the $p = \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{y_i(\mathbf{w}^\top\tilde{\mathbf{x}}_i + b) \le 1-\xi_i\}$, define the indicator function

$$I(\mathbf{x}_i) = \begin{cases} 1, & \text{if } y_i(\mathbf{w}^\top\mathbf{x}_i + b) \le 1-\xi_i \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Then p can be expressed by the following program:

$$p = \sup_{\mathbb{P}} \int_{\mathbb{R}^n} I(\mathbf{x}_i)\mathbb{P}\{\mathbf{x}_i\}\mathrm{d}\mathbf{x}_i \tag{9a}$$

$$\text{s.t.} \int_{\mathbb{R}^n} \mathbb{P}\{\mathbf{x}_i\}\mathrm{d}\mathbf{x}_i = 1 \tag{9b}$$

$$\int_{\mathbb{R}^n} \mathbf{x}_i\mathbb{P}\{\mathbf{x}_i\}\mathrm{d}\mathbf{x}_i = \boldsymbol{\mu}_i \tag{9c}$$

$$\int_{\mathbb{R}^n} \mathbf{x}_i \mathbf{x}_i^\top \mathbb{P}\{\mathbf{x}_i\} \mathrm{d}\mathbf{x}_i = \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top \tag{9d}$$

$$\mathbb{P}\{\mathbf{x}_i\} \geq 0 \tag{9e}$$

The constraints guarantee that $\mathbb{P} \in \mathcal{P}$. The first constraint (9b) is to make sure $\mathbb{P}$ is a probability distribution. The second and third constraints (9c) and (9d) are to guarantee $\mathbb{P}$ has the given first and second moments. Using $z_{0i} \in \mathbb{R}$, $\mathbf{z}_i \in \mathbb{R}^n$ and $\mathbf{Z}_i \in \mathbb{S}^n$ to represent the dual variables of the constraints, then the dual of the above program is:

$$p = \inf_{\mathbf{Z}_i, \mathbf{z}_i, z_{0i}} \ (\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top) \cdot \mathbf{Z}_i + \boldsymbol{\mu}_i^\top \mathbf{z}_i + z_{0i} \tag{10a}$$

$$\text{s.t. } \mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} \geq I(\mathbf{x}_i), \quad \forall \mathbf{x}_i \in \mathbb{R}^n \tag{10b}$$

$$\mathbf{Z}_i \in \mathbb{S}^n, \ \mathbf{z}_i \in \mathbb{R}^n, \ z_{0i} \in \mathbb{R} \tag{10c}$$

where the product $(\cdot)$ is the Frobenius inner product that for matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \cdot \mathbf{B} = \sum_{ij} A_{ij} B_{ij} = \text{Trace}(\mathbf{A}^\top \mathbf{B}) = \text{Trace}(\mathbf{A}\mathbf{B}^\top)$. The strong duality condition guarantees that the optimal values are equal (Isii 1960; Bertsimas and Popescu 2005).

The constraint (10b) can be expressed in two constraints:

$$\mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} \geq 0, \quad \forall \mathbf{x}_i \in \mathbb{R}^n \tag{11a}$$

$$\mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} \geq 1, \quad \text{if } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 1 - \xi_i \tag{11b}$$

Combining the variables $\mathbf{Z}_i, \mathbf{z}_i, z_{0i}$ into one matrix $\mathbf{M}_i$ could obtain:

$$\mathbf{M}_i = \begin{bmatrix} \mathbf{Z}_i & \frac{1}{2}\mathbf{z}_i \\ \frac{1}{2}\mathbf{z}_i^\top & z_{0i} \end{bmatrix} \tag{12}$$

Combining the first and second moments $\boldsymbol{\Sigma}_i, \boldsymbol{\mu}_i$ into one matrix Ω_i:

$$\Omega_i = \begin{bmatrix} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^\top & 1 \end{bmatrix} \tag{13}$$

Then the objective function (10a) becomes $\text{Trace}(\Omega_i \mathbf{M}_i)$. For constraint (11a), it would be $[\mathbf{x}_i^\top 1]\mathbf{M}_i[\mathbf{x}_i^\top 1]^\top \geq 0, \forall \mathbf{x}_i \in \mathbb{R}^n$, i.e. $\mathbf{M}_i \succeq 0$. For constraint (11b), S-lemma is used to reformulate this conditioned constraint.

S-lemma Yakubovich (1971), Pólik and Terlaky (2007) says that let $f, g : \mathbb{R}^n \to \mathbb{R}$ be quadratic functions and suppose that there is an $\bar{x} \in \mathbb{R}^n$ such that $g(\bar{x}) < 0$, then the following two statements are equivalent:

(i) There is no $x \in \mathbb{R}^n$ such that

$$f(x) < 0 \tag{14a}$$

$$g(x) \leq 0 \tag{14b}$$

(ii) There is a nonnegative number $y \geq 0$ such that

$$f(x) + y g(x) \geq 0 \quad \forall x \in \mathbb{R}^n \tag{15}$$

For constraint (11b), it is equivalent to that the following system has no solution $\mathbf{x}_i \in \mathbb{R}^n$ such that

$$\mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} - 1 < 0 \tag{16a}$$

$$y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1 \leq 0 \tag{16b}$$

The first function (16a) is quadratic. The second function (16b) is linear therefore a special case of quadratic functions, and it can achieve the strict negative since it is linear and could obtain all values in $\mathbb{R}$. According to S-lemma, there exists a nonnegative number $\beta_i \geq 0$ such that

$$\mathbf{x}_i^\top \mathbf{Z}_i \mathbf{x}_i + \mathbf{x}_i^\top \mathbf{z}_i + z_{0i} - 1 + \beta_i (y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1) \geq 0 \quad \forall \mathbf{x}_i \in \mathbb{R}^n \tag{17}$$

Then the dual program becomes:

$$p = \inf_{\mathbf{M}_i, \beta_i} \ \text{Trace}(\Omega_i \mathbf{M}_i) \tag{18a}$$

$$\text{s.t. } \mathbf{M}_i \succeq 0, \ \beta_i \geq 0 \tag{18b}$$

$$[\mathbf{x}_i^\top 1]\mathbf{M}_i[\mathbf{x}_i^\top 1]^\top - 1 + \beta_i (y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1) \geq 0 \quad \forall \mathbf{x}_i \in \mathbb{R}^n \tag{18c}$$

The whole program becomes:

$$\min_{\mathbf{w}, b, \xi_i, \mathbf{M}_i, \beta_i} \ \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^m \xi_i \tag{19a}$$

$$\text{s.t. } \mathbf{M}_i \succeq 0, \ \beta_i \geq 0, \ \xi_i \geq 0 \tag{19b}$$

$$\text{Trace}(\Omega_i \mathbf{M}_i) \leq \varepsilon \tag{19c}$$

$$[\mathbf{x}_i^\top 1]\mathbf{M}_i[\mathbf{x}_i^\top 1]^\top - 1 + \beta_i (y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1) \geq 0 \quad \forall \mathbf{x}_i \in \mathbb{R}^n \tag{19d}$$

Since $\mathbf{w}$, b, ξ_i and β_i are all decision variables, it is needed to get rid of the bilinear terms in $\beta_i (y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1)$.

First it could be verified that β_i cannot be zero since $\text{Trace}(\Omega_i \mathbf{M}_i) \leq \varepsilon$, and $0 < \varepsilon < 1$. If $\beta_i = 0$, then (19d) would imply $[\mathbf{x}_i^\top 1]\mathbf{M}_i[\mathbf{x}_i^\top 1]^\top \geq 1 > \varepsilon, \ \forall \mathbf{x}_i \in \mathbb{R}^n$. According to the cyclic property of trace, $\text{Trace}(\mathbf{ABC}) = \text{Trace}(\mathbf{BCA}) = \text{Trace}(\mathbf{CAB})$, therefore, $[\mathbf{x}_i^\top 1]\mathbf{M}_i[\mathbf{x}_i^\top 1]^\top = \text{Trace}([\mathbf{x}_i^\top 1]\mathbf{M}_i[\mathbf{x}_i^\top 1]^\top) = \text{Trace}\left(\begin{bmatrix}\mathbf{x}_i \\ 1\end{bmatrix}[\mathbf{x}_i^\top 1]\mathbf{M}_i\right) \geq 1 > \varepsilon, \ \forall \mathbf{x}_i \in \mathbb{R}^n$. Since $\Omega_i = \begin{bmatrix}\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^\top & 1\end{bmatrix} = \mathbf{E}\left[\begin{bmatrix}\tilde{\mathbf{x}}_i \\ 1\end{bmatrix}[\tilde{\mathbf{x}}_i^\top 1]\right]$, this produces a contradiction. Therefore, $\beta_i > 0$. Then (19d) could be divided by β_i and imply

$$[\mathbf{x}_i^\top 1]\frac{\mathbf{M}_i}{\beta_i}[\mathbf{x}_i^\top 1]^\top - \frac{1}{\beta_i} + (y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1) \geq 0 \quad \forall \mathbf{x}_i \in \mathbb{R}^n \tag{20}$$

For the constraint (19c), since $\varepsilon > 0$, it is equivalent to $\frac{1}{\varepsilon}\text{Trace}(\Omega_i \mathbf{M}_i) - 1 \leq 0$. And since $\beta_i > 0$, it could further get

$$\frac{1}{\varepsilon}\text{Trace}\left(\Omega_i \frac{\mathbf{M}_i}{\beta_i}\right) - \frac{1}{\beta_i} \leq 0 \tag{21}$$

Replace $\frac{\mathbf{M}_i}{\beta_i}$ with $\mathbf{N}_i \succeq 0$, and $\frac{1}{\beta_i}$ with $\alpha_i > 0$, then the two constraints (19c) and (19d) become

$$\frac{1}{\varepsilon}\text{Trace}(\Omega_i \mathbf{N}_i) - \alpha_i \leq 0 \tag{22a}$$

$$[\mathbf{x}_i^\top 1]\mathbf{N}_i[\mathbf{x}_i^\top 1]^\top - \alpha_i + (y_i \mathbf{w}^\top \mathbf{x}_i + y_i b + \xi_i - 1) \geq 0 \quad \forall \mathbf{x}_i \in \mathbb{R}^n \tag{22b}$$

where the second constraint (22b) could further be expressed as a semidefinite constraint as:

$$\mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0 \tag{23}$$

$\alpha_i > 0$ is guaranteed since $\mathbf{N}_i \succeq 0$ and $\frac{1}{\varepsilon}\text{Trace}(\Omega_i \mathbf{N}_i) - \alpha_i \leq 0$.

Then the whole program becomes:

$$\min_{\mathbf{w},b,\xi_i,\mathbf{N}_i,\alpha_i} \ \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \tag{24a}$$

$$\text{s.t. } \alpha_i - \frac{1}{\varepsilon}\text{Trace}(\Omega_i \mathbf{N}_i) \geq 0, \ \ \xi_i \geq 0 \tag{24b}$$

$$\mathbf{N}_i \succeq 0, \ \ \mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0 \tag{24c}$$

This completes the proof. □

The nonlinear term $\|\mathbf{w}\|_2^2$ in the objective function could be replaced by a constraint upper bounding $\|\mathbf{w}\|_2$ by a constant W (Shivaswamy et al. 2006). The adding constraint would be $\|\mathbf{w}\|_2 \leq W$ and the objective function would change to $\sum_{i=1}^{m}\xi_i$. $\|\mathbf{w}\|_2 \leq W$ is a second order cone constraint, and it is contained by semi-definite programs, with a standard SDP form as

$$\|\mathbf{w}\|_2 \leq W \Longleftrightarrow \begin{bmatrix} W\mathbf{I}_n & \mathbf{w} \\ \mathbf{w}^\top & W \end{bmatrix} \succeq 0 \tag{25}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

The standard SDP formulation of **(SVM-SDP)** is

$$\min_{\mathbf{w},b,\xi_i,\mathbf{N}_i,\alpha_i} \ \sum_{i=1}^{m}\xi_i \tag{26a}$$

$$\text{s.t. } \alpha_i - \frac{1}{\varepsilon}\text{Trace}(\Omega_i \mathbf{N}_i) \geq 0, \ \ \xi_i \geq 0, \ \ i = 1, \ldots, m \tag{26b}$$

$$\mathbf{N}_i \succeq 0, \ \ \mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0, \ \ i = 1, \ldots, m \tag{26c}$$

$$\begin{bmatrix} W\mathbf{I}_n & \mathbf{w} \\ \mathbf{w}^\top & W \end{bmatrix} \succeq 0 \tag{26d}$$

This formulation would have the same optimal values for the decision variables with **(SVM-SDP)** when choosing C and W properly.

3.2 Reformulation of (SVM-RCCP) into SOCP

SDP models are generally complicated when computing. SOCP model is a special case of SDP models, but with less variables and more efficient algorithms. The following theorem is to yield SOCP constraints from SDP constraints.

Theorem 2 *The following SDP constraints could yield the following SOCP constraints:*

$$\begin{aligned} &\alpha_i - \frac{1}{\varepsilon}\mathit{Trace}(\Omega_i \mathbf{N}_i) \geq 0, \ \ \mathbf{N}_i \succeq 0, \ \ \mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0 \\ &\Longrightarrow y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\left\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{w}\right\|_2 \end{aligned} \tag{27}$$

Proof First consider the following problem:

$$\inf_{b,\xi_i,\mathbf{N}_i,\alpha_i} \quad y_i b + \xi_i - 1 \tag{28a}$$

$$\text{s.t. } \alpha_i - \frac{1}{\varepsilon}\text{Trace}(\Omega_i \mathbf{N}_i) \geq 0 \tag{28b}$$

$$\mathbf{N}_i \succeq 0 \tag{28c}$$

$$\mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix} \succeq 0 \tag{28d}$$

Let γ_i, $\mathbf{C}_i$, $\overline{\mathbf{D}}_i = \begin{bmatrix} \mathbf{D}_i & \mathbf{d}_i \\ \mathbf{d}_i^\top & d_{0i} \end{bmatrix}$ represent the dual variables of the constraints (28b), (28c), and (28d). Then the Lagrangian is Ghaoui et al. (2003):

$$\begin{aligned}
&\inf_{b,\xi_i,\mathbf{N}_i,\alpha_i} \sup_{\gamma_i\geq 0,\mathbf{C}_i\succeq 0,\overline{\mathbf{D}}_i\succeq 0} \mathscr{L}(\mathbf{w}, b, \xi_i, \mathbf{N}_i, \alpha_i, \gamma_i, \mathbf{C}_i, \overline{\mathbf{D}}_i) \\
&= y_i b + \xi_i - 1 - \gamma_i\Big(\alpha_i - \frac{1}{\varepsilon}\text{Trace}(\Omega_i\mathbf{N}_i)\Big) - \text{Trace}(\mathbf{C}_i\mathbf{N}_i) \\
&\quad - \text{Trace}\Big(\overline{\mathbf{D}}_i, \mathbf{N}_i + \begin{bmatrix} 0 & \frac{1}{2}y_i\mathbf{w} \\ \frac{1}{2}y_i\mathbf{w}^\top & y_i b + \xi_i - 1 - \alpha_i \end{bmatrix}\Big) \\
&= y_i b + \xi_i - 1 - \gamma_i\alpha_i + \frac{\gamma_i}{\varepsilon}\text{Trace}(\Omega_i\mathbf{N}_i) - \text{Trace}(\mathbf{C}_i\mathbf{N}_i) \\
&\quad - \text{Trace}(\overline{\mathbf{D}}_i, \mathbf{N}_i) - y_i\mathbf{w}^\top\mathbf{d}_i - d_{0i}(y_i b + \xi_i - 1 - \alpha_i) \\
&= (y_i b + \xi_i - 1)(1 - d_{0i}) - (\gamma_i - d_{0i})\alpha_i + \text{Trace}\Big(\frac{\gamma_i}{\varepsilon}\Omega_i - \mathbf{C}_i - \overline{\mathbf{D}}_i, \mathbf{N}_i\Big) \\
&\quad - y_i\mathbf{w}^\top\mathbf{d}_i
\end{aligned} \tag{29}$$

This dual function is finite if and only if

$$1 - d_{0i} = 0, \quad \gamma_i - d_{0i} = 0, \quad \frac{\gamma_i}{\varepsilon}\Omega_i - \mathbf{C}_i - \overline{\mathbf{D}}_i = 0 \tag{30}$$

Therefore, $\gamma_i = 1$ and $\frac{1}{\varepsilon}\Omega_i - \overline{\mathbf{D}}_i = \mathbf{C}_i \succeq 0$.

Then the dual problem of (28) is:

$$\sup_{\overline{\mathbf{D}}_i} \quad -y_i\mathbf{w}^\top\mathbf{d}_i \tag{31a}$$

$$\text{s.t. } \frac{1}{\varepsilon}\Omega_i \succeq \overline{\mathbf{D}}_i \succeq 0 \tag{31b}$$

Since $\Omega_i = \begin{bmatrix} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top & \boldsymbol{\mu}_i \\ \boldsymbol{\mu}_i^\top & 1 \end{bmatrix}$, $\overline{\mathbf{D}}_i = \begin{bmatrix} \mathbf{D}_i & \mathbf{d}_i \\ \mathbf{d}_i^\top & d_{0i} \end{bmatrix}$, $d_{0i} = 1$, and $0 < \varepsilon < 1$, the constraint (31b) is equivalent to

$$\begin{bmatrix} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top - \varepsilon\mathbf{D}_i & \boldsymbol{\mu}_i - \varepsilon\mathbf{d}_i \\ \boldsymbol{\mu}_i^\top - \varepsilon\mathbf{d}_i^\top & 1 - \varepsilon \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \varepsilon\mathbf{D}_i & \varepsilon\mathbf{d}_i \\ \varepsilon\mathbf{d}_i^\top & \varepsilon \end{bmatrix} \succeq 0 \tag{32}$$

According to Schur Complement Lemma, for symmetric matrix $\mathbf{S} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{bmatrix}$, where $\mathbf{C}$ is positive definite, then $\mathbf{S} \succeq 0 \Longleftrightarrow \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^\top \succeq 0$. Therefore, (32) is equivalent to

$$\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top - \varepsilon\mathbf{D}_i - \frac{1}{1-\varepsilon}(\boldsymbol{\mu}_i - \varepsilon\mathbf{d}_i)(\boldsymbol{\mu}_i - \varepsilon\mathbf{d}_i)^\top \succeq 0, \quad \varepsilon\mathbf{D}_i - \frac{1}{\varepsilon}\varepsilon\mathbf{d}_i\varepsilon\mathbf{d}_i^\top \succeq 0 \tag{33}$$

i.e.,

$$\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top - \frac{1}{1-\varepsilon}(\boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i)(\boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i)^\top \succeq \varepsilon \mathbf{D}_i \succeq \varepsilon \mathbf{d}_i \mathbf{d}_i^\top \tag{34}$$

The above constraint (34) holds for some $\mathbf{D}_i$ if and only if

$$\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top \succeq \frac{1}{1-\varepsilon}(\boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i)(\boldsymbol{\mu}_i - \varepsilon \mathbf{d}_i)^\top + \varepsilon \mathbf{d}_i \mathbf{d}_i^\top \tag{35}$$

Expand the above constraint

$$\begin{aligned} \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top &\succeq \frac{1}{1-\varepsilon}\boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top - \frac{\varepsilon}{1-\varepsilon}\boldsymbol{\mu}_i \mathbf{d}_i^\top - \frac{\varepsilon}{1-\varepsilon}\mathbf{d}_i \boldsymbol{\mu}_i^\top + \frac{\varepsilon^2}{1-\varepsilon}\mathbf{d}_i \mathbf{d}_i^\top + \varepsilon \mathbf{d}_i \mathbf{d}_i^\top \\ &= \frac{1}{1-\varepsilon}\boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top - \frac{\varepsilon}{1-\varepsilon}(\boldsymbol{\mu}_i \mathbf{d}_i^\top + \mathbf{d}_i \boldsymbol{\mu}_i^\top) + \frac{\varepsilon}{1-\varepsilon}\mathbf{d}_i \mathbf{d}_i^\top \end{aligned} \tag{36}$$

It is equivalent to

$$\boldsymbol{\Sigma}_i \succeq \frac{\varepsilon}{1-\varepsilon}(\boldsymbol{\mu}_i - \mathbf{d}_i)(\boldsymbol{\mu}_i - \mathbf{d}_i)^\top \tag{37}$$

The dual problem (31) becomes

$$\sup_{\mathbf{d}_i} \; -y_i \mathbf{w}^\top \mathbf{d}_i \tag{38a}$$

$$\text{s.t.} \; \frac{1-\varepsilon}{\varepsilon}\boldsymbol{\Sigma}_i - (\boldsymbol{\mu}_i - \mathbf{d}_i)(\boldsymbol{\mu}_i - \mathbf{d}_i)^\top \succeq 0 \tag{38b}$$

From the constraint (38b), we have

$$y_i \mathbf{w}^\top \left(\frac{1-\varepsilon}{\varepsilon}\boldsymbol{\Sigma}_i - (\boldsymbol{\mu}_i - \mathbf{d}_i)(\boldsymbol{\mu}_i - \mathbf{d}_i)^\top \right) y_i \mathbf{w} \geq 0 \tag{39}$$

Since $y_i \in \{+1, -1\}$, we have $y_i^2 = 1$. Then

$$(y_i \mathbf{w}^\top \boldsymbol{\mu}_i - y_i \mathbf{w}^\top \mathbf{d}_i)^2 \leq \frac{1-\varepsilon}{\varepsilon}\mathbf{w}^\top \boldsymbol{\Sigma}_i \mathbf{w} \tag{40}$$

Therefore,

$$-y_i \mathbf{w}^\top \mathbf{d}_i \leq \sqrt{\frac{1-\varepsilon}{\varepsilon}} \left\| \boldsymbol{\Sigma}_i^{\frac{1}{2}} \mathbf{w} \right\|_2 - y_i \mathbf{w}^\top \boldsymbol{\mu}_i \tag{41}$$

In problem (38), $\mathbf{d}_i$ is the only decision variable, therefore, for (41), $\mathbf{d}_i$ is the only variable and the equality could be obtained. The maximum value of $-y_i \mathbf{w}^\top \mathbf{d}_i$ is $\sqrt{\frac{1-\varepsilon}{\varepsilon}} \| \boldsymbol{\Sigma}_i^{\frac{1}{2}} \mathbf{w} \|_2 - y_i \mathbf{w}^\top \boldsymbol{\mu}_i$.

Combine the primal problem (28) and the above result for the dual problem, it could yield that

$$\sqrt{\frac{1-\varepsilon}{\varepsilon}} \left\| \boldsymbol{\Sigma}_i^{\frac{1}{2}} \mathbf{w} \right\|_2 - y_i \mathbf{w}^\top \boldsymbol{\mu}_i \leq y_i b + \xi_i - 1 \tag{42}$$

Or equivalently,

$$y_i (\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}} \left\| \boldsymbol{\Sigma}_i^{\frac{1}{2}} \mathbf{w} \right\|_2 \tag{43}$$

This completes the proof. □

A direct way to reformulate **(SVM-RCCP)** into SDP model is to use multivariate Chebyshev inequality (Bhattacharyya et al. 2004; Shivaswamy et al. 2006). Let $\tilde{\mathbf{x}} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote random vector $\tilde{\mathbf{x}}$ with mean $\boldsymbol{\mu}$ and convariance matrix $\boldsymbol{\Sigma}$, the multivariate Chebyshev inequality (Marshall and Olkin 1960; Bertsimas and Popescu 2005) states that for an arbitrary closed convex set S, the supremum of the probability that $\tilde{\mathbf{x}}$ takes a value in S is

$$\sup_{\tilde{\mathbf{x}} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})} \mathbb{P}\{\tilde{\mathbf{x}} \in S\} = \frac{1}{1+d^2} \tag{44a}$$

$$d^2 = \inf_{\mathbf{x} \in S} (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \tag{44b}$$

For SVM constraint, the $S = \{y(\mathbf{w}^\top \mathbf{x} + b) \leq 1-\xi\}$ is a half-space produced by a hyperplane and therefore a closed convex set, using the above inequality we could obtain (Lanckriet et al. 2002)

$$\sup_{\tilde{\mathbf{x}} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})} \mathbb{P}\Big\{ y(\mathbf{w}^\top \tilde{\mathbf{x}} + b) \leq 1-\xi \Big\} = \frac{1}{1+d^2} \tag{45}$$

where $d^2 = \inf_{y(\mathbf{w}^\top \mathbf{x}+b) \leq 1-\xi} (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})$.

When $y(\mathbf{w}^\top \boldsymbol{\mu}+b) \leq 1-\xi$, then take $\mathbf{x} = \boldsymbol{\mu}$ could get $d^2 = 0$ and $\sup_{\tilde{\mathbf{x}} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})} \mathbb{P}\{y(\mathbf{w}^\top \tilde{\mathbf{x}} + b) \leq 1-\xi\} = 1$, which is trivial and this condition would be invalid when the bound $\varepsilon < 1$.

When $y(\mathbf{w}^\top \boldsymbol{\mu}+b) > 1-\xi$, let $\mathbf{u} = \Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$, $\mathbf{v} = y\boldsymbol{\Sigma}^{1/2}\mathbf{w}$ and $\gamma = -y(\mathbf{w}^\top \boldsymbol{\mu}+b) + 1-\xi < 0$, then $d^2 = \inf_{\mathbf{v}^\top \mathbf{u} \leq \gamma} \mathbf{u}^\top \mathbf{u}$. And the Lagrangian is $L(\mathbf{u}, \lambda) = \mathbf{u}^\top \mathbf{u} + \lambda(\mathbf{v}^\top \mathbf{u} - \gamma)$ with $\lambda \geq 0$. Take the derivative to be zero, at the optimum, $2\mathbf{u} = \lambda \mathbf{v}$ and $\mathbf{v}^\top \mathbf{u} = \gamma$. Therefore,

$$\begin{aligned} d^2 &= \inf_{y(\mathbf{w}^\top \mathbf{x}+b) \leq 1-\xi} (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) = \mathbf{u}^\top \mathbf{u} = \frac{\gamma^2}{\mathbf{v}^\top \mathbf{v}} \\ &= \frac{\big(y(\mathbf{w}^\top \boldsymbol{\mu} + b) - 1 + \xi\big)^2}{\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}} \end{aligned} \tag{46}$$

For $\sup_{\tilde{\mathbf{x}} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})} \mathbb{P}\{y(\mathbf{w}^\top \tilde{\mathbf{x}} + b) \leq 1-\xi\} \leq \varepsilon$, it is equivalent to $1/(1+d^2) \leq \varepsilon$, or $d^2 \geq (1-\varepsilon)/\varepsilon$. Since $y(\mathbf{w}^\top \boldsymbol{\mu} + b) > 1-\xi$, i.e. $y(\mathbf{w}^\top \boldsymbol{\mu} + b) - 1 + \xi > 0$, therefore

$$y(\mathbf{w}^\top \boldsymbol{\mu} + b) - 1 + \xi \geq \sqrt{\frac{1-\varepsilon}{\varepsilon}} \left\| \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{w} \right\|_2 \tag{47}$$

Applying the above result to **(SVM-RCCP)**, the Chebyshev based reformulation could be achieved utilizing the mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ of each uncertain training point $\tilde{\mathbf{x}}_i$ as the following robust model (Bhattacharyya et al. 2004; Shivaswamy et al. 2006):

(SVM − SOCP)

$$\min_{\mathbf{w}, b, \xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^{m} \xi_i \tag{48a}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}} \left\| \boldsymbol{\Sigma}_i^{\frac{1}{2}} \mathbf{w} \right\|_2 \tag{48b}$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m \tag{48c}$$

3.3 The geometric interpretation of (SVM-SOCP)

The geometric interpretation of the SOCP constraint is that, for each point $\mathbf{x}_i$, it is no longer a single point, but an ellipsoid centered at $\boldsymbol{\mu}_i$, and shaped with the covariance matrix $\sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}$:

$$\mathscr{E}\left(\boldsymbol{\mu}_i, \sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\right) = \left\{\mathbf{x} = \boldsymbol{\mu}_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{a} : \ \|\mathbf{a}\|_2 \leq 1\right\} \tag{49}$$

The SOCP constraint (48b) is satisfied if and only if

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall \mathbf{x}_i \in \mathscr{E}\left(\boldsymbol{\mu}_i, \sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\right) \tag{50}$$

Therefore, this constraint is defining an uncertainty set $\mathscr{E}\left(\boldsymbol{\mu}_i, \sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}\right)$ for each uncertain training data point $\mathbf{x}_i$. If all the points in the uncertainty set satisfy $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$, then the chance-constraint is guaranteed to be satisfied. This transforms the RCCP into a robust optimization problem over the uncertainty set.

Figure 1 shows how the SVM works when the fixed data points are replaced with their corresponding ellipsoid uncertainty sets. Figure 1a is the original data. The red line is the separating line $\mathbf{w}^\top \mathbf{x} + b = 0$. The blue and green dash lines are the lines passing through the support vectors, i.e., the lines $\mathbf{w}^\top \mathbf{x} + b = \pm 1$. As shown in the plot, the margin width,

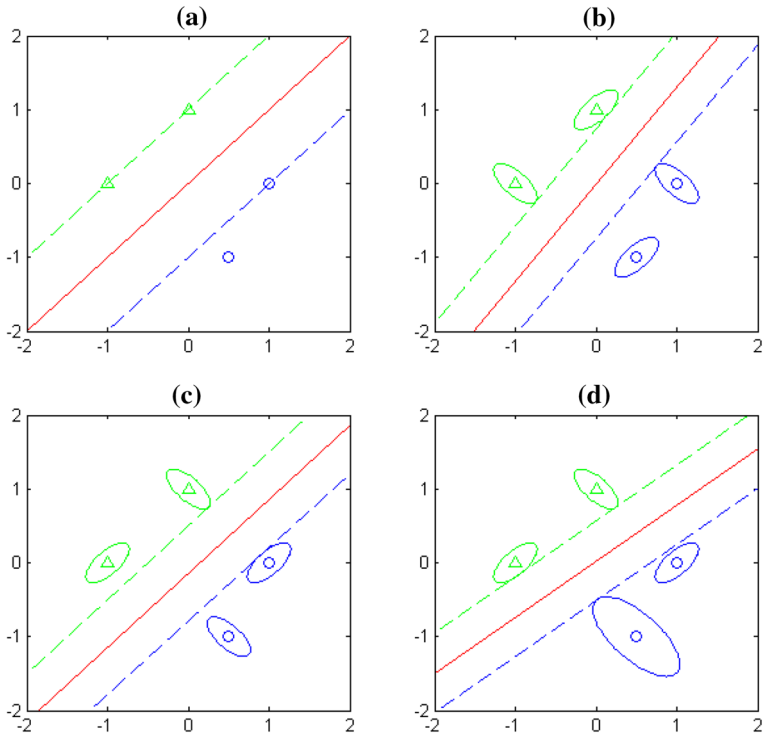

Fig. 1 Geometric interpretation of SVM classfier for data with different uncertainty sets

i.e., the distance between these two dash lines are maximized on the data. It is worth noting that not all points are support vectors in this figure. In Fig. 1b, the data are enclosed into ellipsoid uncertainty sets, the lines are the same meaning as Fig. 1a. So instead of making the single points to be on the right side of the margin, the dash lines should make the whole ellipsoids to be beyond the margin. From this plot, it could be seen that the green dash line are forced to be tangent to the two green ellipsoids from the inner side. But this is not always the case. Figure 1c shows a case when both dash lines are only tangent to one ellipsoid each, but still obtain the maximum margin classifier. Figure 1d shows how the size of the uncertainty set would affect the classifier. All the other settings are the same with Fig. 1c, except that the lower blue point now has a larger uncertainty set. To make this uncertainty set satisfy the constraint, the separating lines are pushed up as shown in the plot, and the upper blue ellipsoid is not touching the margin line any more.

4 Estimation errors and performance measures

In practice, the distribution properties are often unknown but need to be estimated from data. For example, if an uncertain data point $\tilde{\mathbf{x}}_i = [\tilde{x}_{i1}, \ldots, \tilde{x}_{in}]^\top$ has N samples $\mathbf{x}_{i_k}, k = 1, \ldots, N$, then the sample mean $\bar{\mathbf{x}}_i = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_{i_k}$ is used to estimate the mean vector $\boldsymbol{\mu}_i = \mathbf{E}[\tilde{\mathbf{x}}_i]$, and the sample covariance $\mathbf{S}_i = \frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{x}_{i_k} - \bar{\mathbf{x}}_i)(\mathbf{x}_{i_k} - \bar{\mathbf{x}}_i)^\top$ is used to estimate the covariance matrix $\boldsymbol{\Sigma}_i = \mathbf{E}\big[(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_i)(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_i)^\top\big]$. However, these could cause possible estimation errors. For example, the sample mean $\bar{\mathbf{x}}_i$ itself is a random vector, with mean equal to $\boldsymbol{\mu}_i$, and the variance of the jth element is equal to σ_{ij}^2/N, where σ_{ij}^2 is the variance of the random variable $\tilde{x}_{ij}$ and unknown. Three special cases when the mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ are not exactly known are discussed here.

The first case is when $\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$, and $\boldsymbol{\Sigma}_i = \mathbf{S}_i$ does not have variance. The interval of $\boldsymbol{\mu}_i$ actually works for each element in the vector, i.e. $\mu_{ij} \in [\mu_{ij}^-, \mu_{ij}^+]$, $j = 1, \ldots, n$. As a robust reformulation of the **(SVM-SOCP)** model, the constraint (48b) becomes $\sum_j \min(y_i\mu_{ij}^- w_j, y_i\mu_{ij}^+ w_j) + y_i b \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{w}\|_2$, and the whole model could be written as:

$$\textbf{(SVM} - \textbf{SOCP} - \textbf{Mu1)}$$

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \tag{51a}$$

$$\text{s.t.} \quad \sum_j z_{ij} + y_i b \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\left\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{w}\right\|_2 \tag{51b}$$

$$z_{ij} \leq y_i\mu_{ij}^- w_j, \; z_{ij} \leq y_i\mu_{ij}^+ w_j \tag{51c}$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m \tag{51d}$$

This case could be applied when the confidence interval of μ_{ij} could be estimated. For a random variable $\tilde{x}_{ij}$ with normal distribution, and N samples $x_{ij_k}, k = 1, \ldots, N$, the sample mean $\bar{x}_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_{ij_k}$, the unbiased sample variance $s_{ij}^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ij_k} - \bar{x}_{ij})^2$, then $\frac{\bar{x}_{ij}-\mu_{ij}}{s_{ij}/\sqrt{N}} \sim t_{N-1}$. The confidence interval of μ_{ij} is $[\bar{x}_{ij} - t_{crit}\cdot s_{ij}/\sqrt{N}, \bar{x}_{ij} + t_{crit}\cdot s_{ij}/\sqrt{N}]$, where t_{crit} is the coefficient corresponding to the confidence level $1-\alpha$ and the degree of freedom $N-1$. For n-dimensional vector $\tilde{\mathbf{x}}_i \in \mathbb{R}^n$, the Bonferroni correction factor uses α/n

instead of α for each of the n univariate confidence interval. The geometric interpretation of this case is that, for each point $\mathbf{x}_i$, it is replaced by a union of ellipsoids, with center varies in the hyper-rectangle $[\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$, and shaped with the covariance matrix $\sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}$.

The second case is when $(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i) \leq \nu_i^2$, and $\boldsymbol{\Sigma}_i = \mathbf{S}_i$ has no variance. Since the Bonferroni correction is for the case when the random variables $\tilde{x}_{ij}$ in $\tilde{\mathbf{x}}_i$ are independent, it would over-correct and result in lower α and larger robust region than it needs to be when they are not independent. The Hotelling's T-square test statistic $T^2 = N(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$, and it has the property that $\frac{N-n}{n(N-1)}T^2 \sim F(n, N-n)$. Then the confidence region for $\boldsymbol{\mu}_i$ is $T^2 \leq \frac{n(N-1)}{N-n}F_{crit}$, i.e., $(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)^\top \mathbf{S}_i^{-1}(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i) \leq \frac{n(N-1)}{N(N-n)}F_{crit}$, where F_{crit} is the coefficient corresponding to the confidence level $1-\alpha$ and the degree of freedom $(n, N-n)$. Let $\nu_i^2 = \frac{n(N-1)}{N(N-n)}F_{crit}$, the geometric interpretation is that, the mean vector $\boldsymbol{\mu}_i$ varies in an ellipsoid centered at $\bar{\mathbf{x}}_i$ and shaped with $\nu_i \boldsymbol{\Sigma}_i^{\frac{1}{2}}$. Then the uncertainty set for each point $\mathbf{x}_i$ is a union of ellipsoids, with center varies in the ellipsoid $\mathscr{E}\left(\bar{\mathbf{x}}_i, \nu_i \boldsymbol{\Sigma}_i^{\frac{1}{2}}\right)$, and shaped with $\sqrt{\frac{1-\varepsilon}{\varepsilon}}\boldsymbol{\Sigma}_i^{\frac{1}{2}}$. This has a more concise form that the union of these ellipsoids is also an ellipsoid $\mathscr{E}(\bar{\mathbf{x}}_i, \left(\sqrt{\frac{1-\varepsilon}{\varepsilon}} + \nu_i\right)\boldsymbol{\Sigma}_i^{\frac{1}{2}})$. Lanckriet et al. (2002) also proved this mathematically. The model is then:

$$\textbf{(SVM} - \textbf{SOCP} - \textbf{Mu2)}$$

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \tag{52a}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq 1 - \xi_i + \left(\sqrt{\frac{1-\varepsilon}{\varepsilon}} + \nu_i\right)\left\|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\mathbf{w}\right\|_2 \tag{52b}$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m \tag{52c}$$

The third case is when $\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i$ has no variance, but the covariance matrix has estimation uncertainty $\|\boldsymbol{\Sigma}_i - \mathbf{S}_i\|_F \leq \rho_i$, where the matrix norm is the Frobenius norm $\|\mathbf{A}\|_F^2 = \text{Trace}(\mathbf{A}^\top\mathbf{A}) = \sum_{ij} A_{ij}^2$. Lanckriet et al. (2002) proved that in this case, the uncertainty set becomes $\mathscr{E}\left(\boldsymbol{\mu}_i, \sqrt{\frac{1-\varepsilon}{\varepsilon}}\left(\boldsymbol{\Sigma}_i + \rho_i I_n\right)^{\frac{1}{2}}\right)$. And the model is:

$$\textbf{(SVM} - \textbf{SOCP} - \textbf{Cov)}$$

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \tag{53a}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq 1 - \xi_i + \sqrt{\frac{1-\varepsilon}{\varepsilon}}\left\|\left(\boldsymbol{\Sigma}_i + \rho_i I_n\right)^{\frac{1}{2}}\mathbf{w}\right\|_2 \tag{53b}$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m \tag{53c}$$

The geometric interpretations of the three cases are shown in Fig. 2. Figure 2a is when $\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$. The blue ellipsoid is the original robust region. The green solid rectangle is the area that $\boldsymbol{\mu}_i$ can be varied. The green dash ellipsoids are the robust regions when $\boldsymbol{\mu}_i$ varies in $[\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$. And the big blue boundary is the resulting robust region of this case. Figure 2b is when $(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i) \leq \nu_i^2$. The small blue ellipsoid is the original robust region. The green solid ellipsoid and green dash ellipsoids have similar meaning as in Fig. 2a. The resulting robust region is the big blue ellipsoid. Figure 2c is when $\|\boldsymbol{\Sigma}_i - \mathbf{S}_i\|_F \leq \rho_i$. The

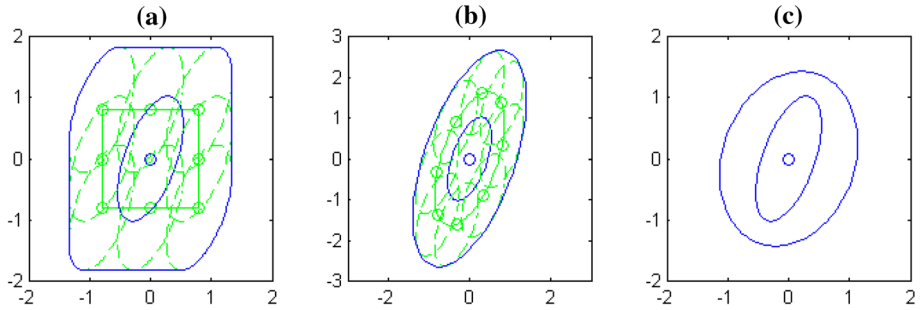

Fig. 2 Geometric interpetation of different estimation errors

inner blue ellipsoid and outer ellipsoid represent the original and resulting robust region, respectively. In this case, the shape of the ellipsoids changed because the matrices changed for the ellipsoids.

Since now the data points are uncertain, the performance measures are worth discussed. One direct way is to use the Test Set Accuracy (TSA) which is computed by counting the number of correctly predicted labels in the test data set and divided by the size of the test set. The class label y_i is decided by the sign$(\mathbf{w}^\top \mathbf{x}_i + b)$. When there are replicates $\mathbf{x}_{i_k}$ for the test point $\mathbf{x}_i$, the class label y_i is decided by the majority label of the replicates sign$(\mathbf{w}^\top \mathbf{x}_{i_k} + b)$.

Another way is proposed by Ben-Tal et al. (2011) to use the nominal error and optimal error. The nominal error is similar to TSA but the opposite, i.e., TSA + NomErr = 1. The expression for NomErr is:

$$\text{NomErr} = \frac{\sum_i 1_{y_i^{pr} \neq y_i}}{\#\text{ test datapoints}} \times 100\,\% \tag{54}$$

The optimal error is calculated based on the probability of misclassification. For $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \leq 0\} \leq \varepsilon$, it can be similarly transformed into $y_i(\mathbf{w}^\top \boldsymbol{\mu}_i + b) \geq \sqrt{\frac{1-\varepsilon}{\varepsilon}} \|\boldsymbol{\Sigma}_i^{\frac{1}{2}} \mathbf{w}\|_2$. And this could derive that the least value of ε is

$$\varepsilon_{opt} = \frac{\mathbf{w}^\top \boldsymbol{\Sigma}_i \mathbf{w}}{(\mathbf{w}^\top \boldsymbol{\mu}_i + b)^2 + \mathbf{w}^\top \boldsymbol{\Sigma}_i \mathbf{w}} \tag{55}$$

Then the OptErr of the data point $\mathbf{x}_i$ is

$$\text{OptErr}_i = \begin{cases} 1, & \text{if } y_i^{pr} \neq y_i \\ \varepsilon_{opt}, & \text{if } y_i^{pr} = y_i \end{cases} \tag{56}$$

And the OptErr of the whole test set is

$$\text{OptErr} = \frac{\sum_i \text{OptErr}_i}{\#\text{ test datapoints}} \times 100\,\% \tag{57}$$

5 Numerical experiments

5.1 The equivalence of (SVM-SDP) and (SVM-SOCP)

The model **(SVM-SDP)** and model **(SVM-SOCP)** are equivalent since they both use the exact supremum of the chance constraints $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \leq 1 - \xi_i\}$ and both based

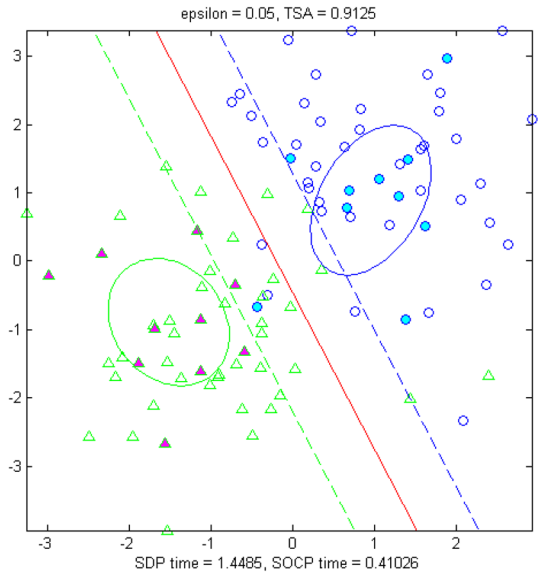

Fig. 3 Sythetic data classification result

on the exact means and covariance matrices of the random data points. Multiple numerical experiments in MATLAB using SeDuMi solver for both models on YALMIP platform also show that these two formulations would obtain the same result.

In numerical experiments, to make the results easier to interpret, firstly, synthetic data is generated using 2-dimensional normal distribution with the $+1$ class generated by normal distribution with $\boldsymbol{\mu}_+ = [1, 1]^\top$, $\boldsymbol{\Sigma}_+ = \mathbf{I}$ the identity matrix, and -1 class generated by normal distribution with $\boldsymbol{\mu}_- = [-1, -1]^\top$, $\boldsymbol{\Sigma}_- = \mathbf{I}$, each class has 50 points. Then for each class, 10 points are randomly picked as the training points, the rest are the test points. We need to get the $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ for each training point. In our experiment, $\boldsymbol{\mu}_i$ is set to be the current value of the point, $\boldsymbol{\Sigma}_i$ is calculated based on the convariance matrix of the training points for each class. For example, for the $+1$ class, we have 10 training points, then we calculated the covariance matrix of these 10 points, and multiplied by 0.01 to shrink the area that each points could move. This shrinking effect is to make the ellipsoid size to be 1/10 of the original covariance matrix ellipsoid and is reasonable as an uncertainty set for each data point. With these values, we tested both models and get Fig. 3.

In Fig. 3, the blue circles are the points of the $+1$ class with the ones filled inside with cyan color indicating the training points, and the rest are test points. The green triangles are the points of the -1 class with the ones filled inside with magenta color indicating the training points, and the rest are test points. The blue ellipsoid is drawn based on the training points mean and covariance matrix of the $+1$ class with the expression $\mathscr{E}\big(\boldsymbol{\mu}_+, \boldsymbol{\Sigma}_+^{\frac{1}{2}}\big) = \{\mathbf{x} = \boldsymbol{\mu}_+ + \boldsymbol{\Sigma}_+^{\frac{1}{2}}\mathbf{a} : \ \|\mathbf{a}\|_2 \leq 1\}$. The green ellipsoid is drawn similarly for the -1 class. These two ellipsoid could give us some direct impression for the distribution of the training samples. And to shrink the ellipsoid size to be the uncertainty set for each training point is much more reasonable than to directly use these original big ellipsoids shown in the figure. The red solid line is the separating line $\mathbf{w}^\top\mathbf{x} + b = 0$ calculated by both the SDP and SOCP models. The blue and green dash lines are the margin lines, i.e., the lines $\mathbf{w}^\top\mathbf{x} + b = \pm 1$.

Figure 3 shows only one red line, one blue dash line, and one green dash line. In fact, the separating and margin lines calculated by both models are drawn in this plot. One line means

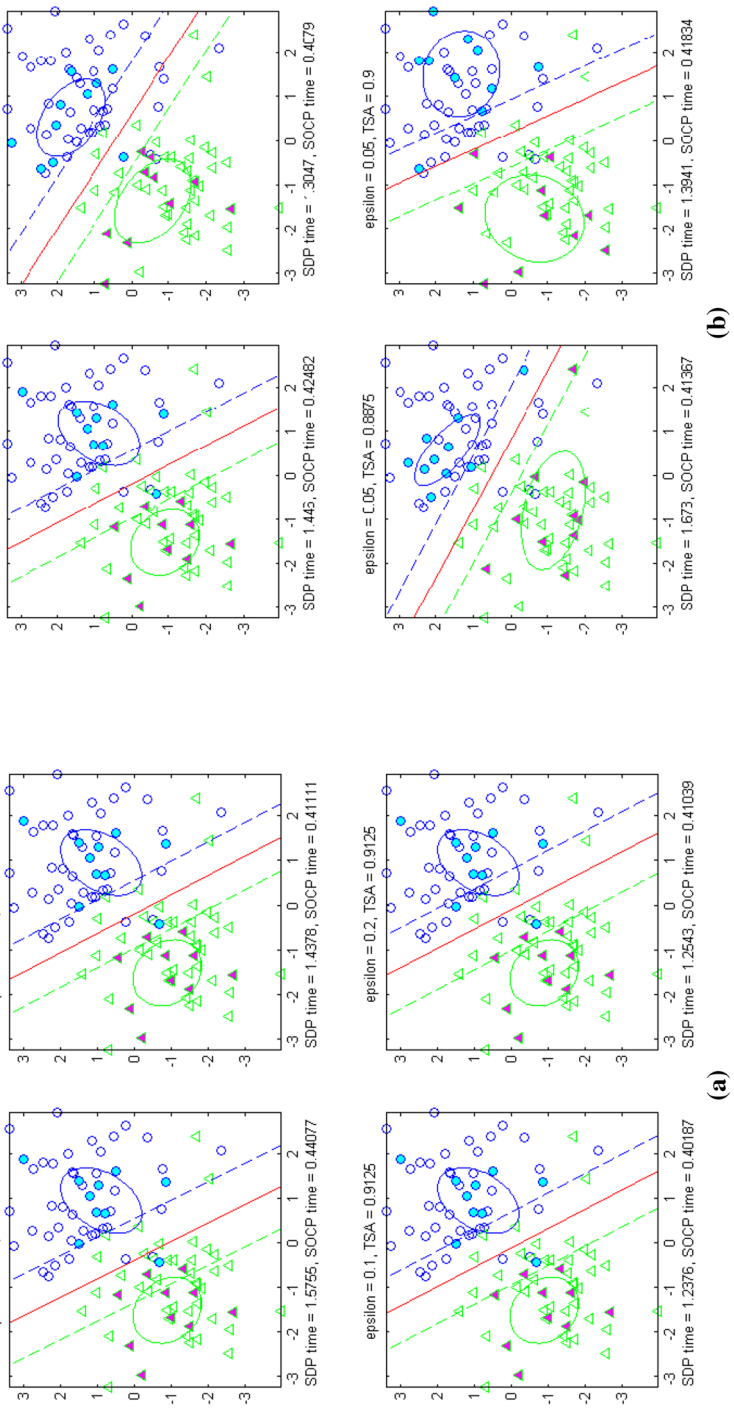

Fig. 4 Sythetic data classification result with different ε and different training set

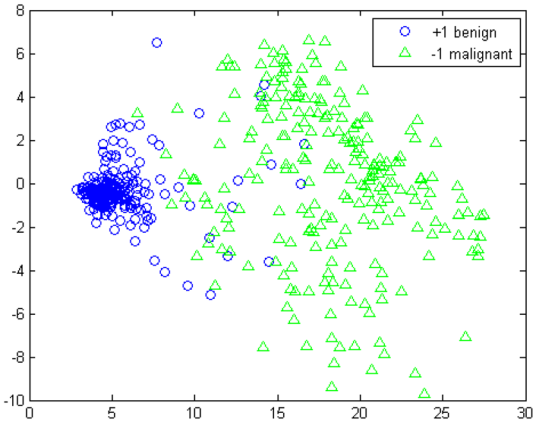

Fig. 5 First two principle components of Wisconsin breast cancer data

the two models are getting exactly the same result. The running time for both models are also shown in this figure. As SDP models have much more variables than the SOCP model, and the intrinsic complexity than the SOCP model, the SDP took longer time than the SOCP model. The TSA for this experiment is 91.25 %.

To compare how the chance constraint probability controller ε would affect the performance of the models, we fix the training and test point. By changing only the value of ε, we have Fig. 4a. It shows that the SDP model and SOCP model are getting the same result no matter how ε changes, while the SOCP model is more efficient than the SDP model. Generally, small ε values would need a longer running time. The test-set accuracy does not have common trend with ε when different data are generated randomly in different runs. And it highly depends on the split of the training set and test set as shown in Fig. 4b. It shows that the SDP model and SOCP models are still getting the same result even though the training points are changing between runs. Besides this, another important aspect of this figure is that with the same data set, how the split of the training set and test set would dramatically change the separation line. To test the equivalence of the two models, we also used $\boldsymbol{\Sigma}_i = 0.01\mathbf{I}$ for the synthetic data, and the results from the two models are still the same.

Besides the synthetic data, we also tested on real data, the Wisconsin breast cancer data from UCI dataset. This data contains 699 samples, while 16 samples have missing values so we do not use, resulting in 683 samples. Among these samples, 444 are benign, we record as $+1$ class; 239 are malignant, we record as -1 class. These samples each has 10 attributes, with the first attribute to be the sample id number, which we do not include into the features. This results in 9-dimensional features.

To be able to show in figure how the data is distributed, firstly, we use PCA to extract the first two principle components and get the 2-dimensional data plot in Fig. 5. The -1 malignant class actually has fewer data points than the $+1$ benign class but the -1 points are more spread.

The average results for Wisconsin breast cancer data over 20 runs with random partitions are shown in Table 1. The first subtable is when 20 % of the data are used as training and the remaining 80 % are test data; the second subtable is when 80 % are training and the remaining 20 % are test. The boxplots of the results are shown in Fig. 6. Both SDP and SOCP models are getting the same result, while SOCP model runs more efficiently than SDP model. When there are more training points, the TSA is higher, but the training time also increases.

Table 1 Wisconsin breast cancer data with different training and test partitions

	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
20 % training, 80 % test				
Test set accuracy (%)	96.8 ± 0.6	96.4 ± 1.0	96.1 ± 1.1	96.1 ± 1.2
SDP running time	29.0 ± 2.5	26.5 ± 3.2	23.1 ± 3.2	21.1 ± 2.9
SOCP running time	1.3 ± 0.2	1.2 ± 0.3	1.2 ± 0.3	1.2 ± 0.3
80 % training, 20 % test				
Test set accuracy (%)	97.4 ± 0.9	97.1 ± 1.0	97.0 ± 1.1	97.1 ± 1.0
SDP running time	155.6 ± 9.7	141.8 ± 9.1	134.2 ± 13.7	121.4 ± 11.7
SOCP running time	3.3 ± 0.1	4.1 ± 0.2	5.0 ± 0.4	5.8 ± 0.6

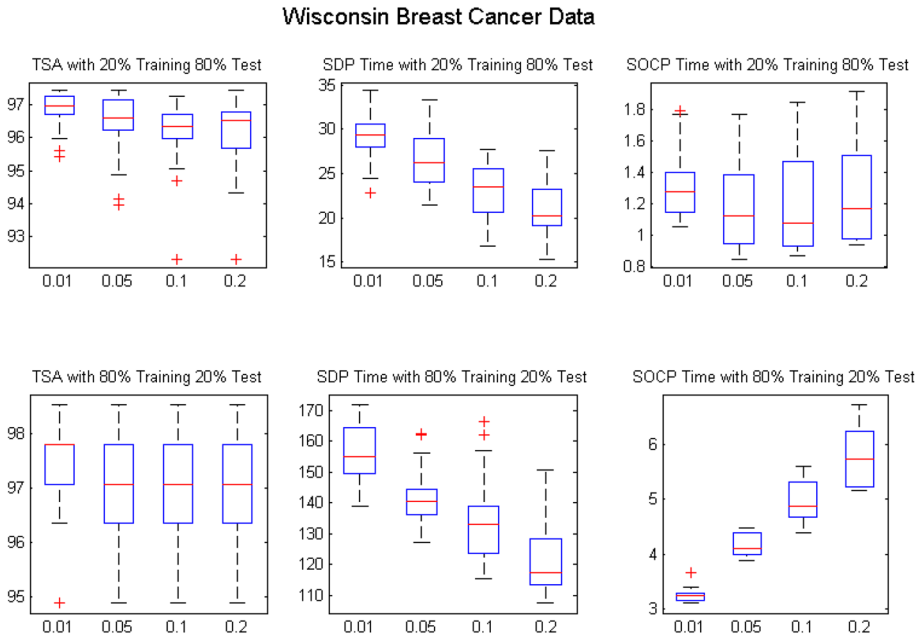

Fig. 6 Boxplots of Wisconsin breast cancer data with different training and test partitions

Besides the Wisconsin breast cancer data, Ionosphere data is also used in the experiments. Ionosphere data is 34-dimensional data, with 225 samples for the $+1$ good class, and 126 samples for the -1 bad class. The first two principle components plot is shown in Fig. 7. It could be seen that the two classes of data are overlapping with each other.

Since 34 dimensions are time consuming when computing, we used PCA to extract the first 15 dimensions, then performed both SDP and SOCP models on the 15-dimensional data. The results for the extracted Ionosphere data over 20 runs with random partitions are shown in Table 2 and Fig. 8. When $\varepsilon = 0.01$, the robust region is too large that the resulting $\mathbf{w}$ will be approximately $\mathbf{0}$ so the models are not getting meaningful results. Therefore, we used $\varepsilon = 0.02$ instead. From these two tables, SDP and SOCP models are still getting the same result, while SOCP model runs in less time. The TSA of Ionosphere data is less than the breast

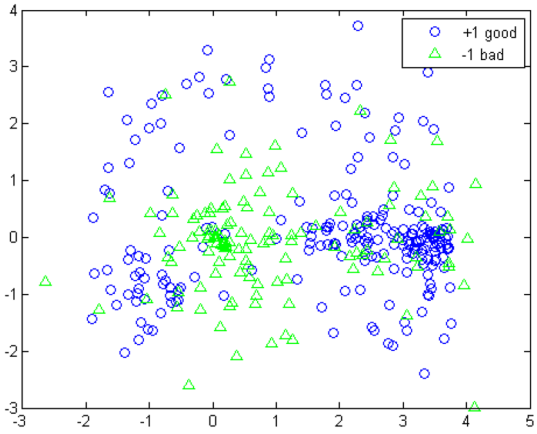

Fig. 7 First two principle components of ionosphere data

Table 2 Extracted ionosphere data with different training and test partitions

	$\varepsilon = 0.02$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
20 % training, 80 % test				
Test set accuracy (%)	84.0 ± 2.5	84.4 ± 2.1	84.1 ± 2.2	84.2 ± 2.2
SDP running time	20.6 ± 1.8	18.3 ± 1.6	18.1 ± 2.1	19.1 ± 2.4
SOCP running time	1.1 ± 0.2	1.1 ± 0.3	1.0 ± 0.3	1.0 ± 0.4
80 % training, 20 % test				
Test set accuracy (%)	86.9 ± 3.6	87.8 ± 3.8	87.2 ± 3.9	87.2 ± 4.3
SDP running time	107.4 ± 7.6	97.9 ± 7.3	96.0 ± 10.0	95.7 ± 8.3
SOCP running time	2.4 ± 0.2	3.0 ± 0.4	3.7 ± 0.5	4.6 ± 0.3

cancer data. This is reasonable seeing from the first two principle components plot of the two datasets.

5.2 Estimation and performance issues

Considering the estimation errors discussed in Sect. 4, the three cases $\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$, $(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i) \leq \nu_i^2$, and $\|\boldsymbol{\Sigma}_i - \mathbf{S}_i\|_F \leq \rho_i$ were experimented on the tow norm data, the Wisconsin breast cancer data, and the Ionosphere data. For each data point $\mathbf{x}_i$, $N = 50$ replicates $\mathbf{x}_{i_k} (k = 1, \ldots, N)$ were generated with mean equal to the value of the data point $\mathbf{x}_i$, and covariance equal to 0.01 times the covariance of the training dataset. For the two norm data, since we generated the data using $\boldsymbol{\Sigma}_+ = \boldsymbol{\Sigma}_- = \mathbf{I}$, the replicates generation covariance used $0.01\mathbf{I}$.

Then for each data point $\mathbf{x}_i$ with 50 samples $\mathbf{x}_{i_k}$, the sample mean $\bar{\mathbf{x}}_i = \frac{1}{N}\sum_{k=1}^N \mathbf{x}_{i_k}$ and the sample covariance $\mathbf{S}_i = \frac{1}{N-1}\sum_{k=1}^N (\mathbf{x}_{i_k} - \bar{\mathbf{x}}_i)(\mathbf{x}_{i_k} - \bar{\mathbf{x}}_i)^\top$ are calculated to estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. When estimation errors are considered, the first case $\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$, the confidence interval of μ_{ij} is $[\bar{x}_{ij} - t_{crit} \cdot s_{ij}/\sqrt{N}, \bar{x}_{ij} + t_{crit} \cdot s_{ij}/\sqrt{N}]$. We used $\alpha = 0.1$ in our experiment, since $\tilde{\mathbf{x}}_i \in \mathbb{R}^n$, the Bonferroni correction factor requires α/n for each of the n univariate confidence interval. Therefore, for the two norm data $\tilde{\mathbf{x}}_i \in \mathbb{R}^2$, the t_{crit} has confidence level

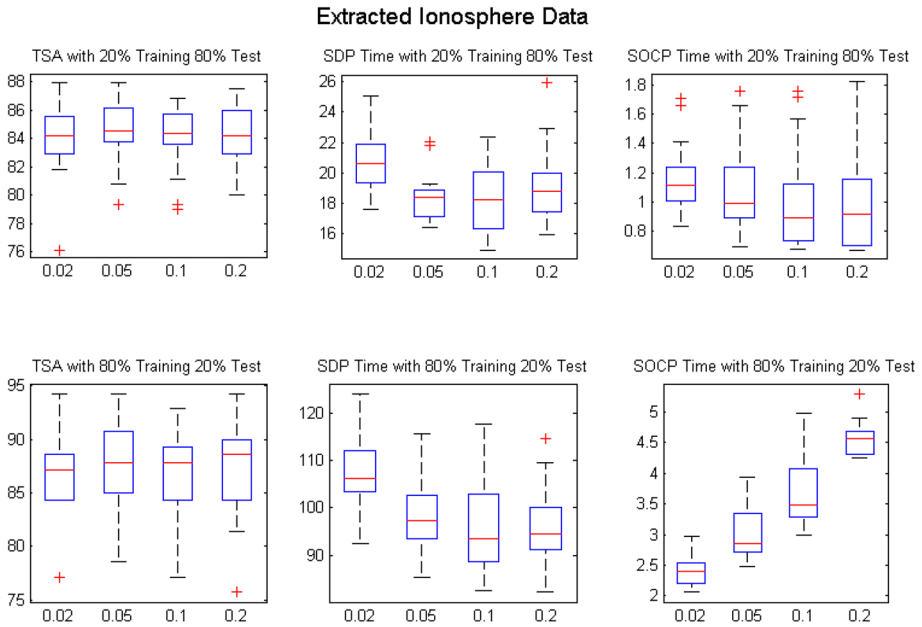

Fig. 8 Boxplots of extracted ionosphere data with different training and test partitions

Table 3 Performance measure results considering estimation errors

	$\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i\ \boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i \in \mathscr{E}(\bar{\mathbf{x}}_i, \nu_i \boldsymbol{\Sigma}_i^{\frac{1}{2}})$ $\boldsymbol{\Sigma}_i = \mathbf{S}_i$	$\boldsymbol{\mu}_i = \bar{\mathbf{x}}_i \Vert \boldsymbol{\Sigma}_i - \mathbf{S}_i \Vert_F \leq \rho_i$
Two norm				
NomErr (%)	9.75 ± 3.69	9.63 ± 3.42	9.63 ± 3.61	9.69 ± 3.82
OptErr (%)	12.44 ± 3.92	12.62 ± 3.92	12.46 ± 3.84	12.43 ± 3.75
Time	0.59 ± 0.10	0.74 ± 0.09	0.59 ± 0.09	0.60 ± 0.08
Breast cancer				
NomErr (%)	4.16 ± 1.03	3.99 ± 0.83	4.07 ± 0.92	3.66 ± 0.82
OptErr (%)	6.82 ± 0.99	6.61 ± 0.88	6.64 ± 0.89	6.26 ± 0.75
Time	1.94 ± 0.24	6.18 ± 0.77	2.01 ± 0.26	2.10 ± 0.28
Ionosphere data				
NomErr (%)	15.21 ± 2.58	15.60 ± 2.17	15.55 ± 2.41	19.15 ± 6.81
OptErr (%)	18.87 ± 2.20	18.92 ± 2.15	18.93 ± 2.21	22.02 ± 5.76
Time	1.35 ± 0.13	5.98 ± 0.69	1.38 ± 0.16	1.37 ± 0.18

of $1-0.05$; for the breast cancer data $\tilde{\mathbf{x}}_i \in \mathbb{R}^9$, the t_{crit} has confidence level of $1-0.0111$; for the extracted Ionosphere data $\tilde{\mathbf{x}}_i \in \mathbb{R}^{15}$, the t_{crit} has confidence level of $1-0.0067$. And all t_{crit}s have the degree of freedom $N-1=49$. Model (**SVM-SOCP-Mu1**) was used to calculate this case.

For the second case $(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_i - \bar{\mathbf{x}}_i) \leq \nu_i^2$, as discussed in Sect. 4, $\nu_i^2 = \frac{n(N-1)}{N(N-n)} F_{crit}$. Since $\alpha = 0.1$, all F_{crit}s have the confidence level of $1-0.1$. The degree of

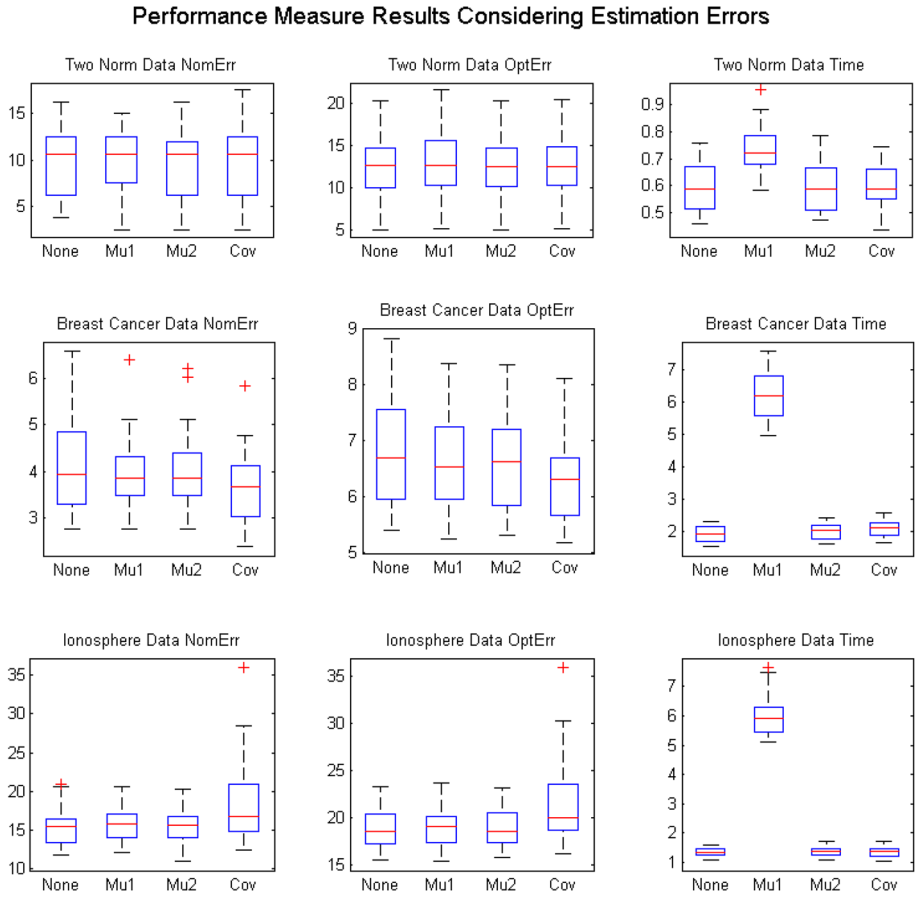

Fig. 9 Boxplots of performance measure results considering estimation errors

freedom for the two norm data is $(2, 48)$, for the breast cancer data is $(9, 41)$, and for the breast cancer data is $(15, 35)$. Model **(SVM-SOCP-Mu2)** was used to calculate this case. For the third case $\|\boldsymbol{\Sigma}_i - \mathbf{S}_i\|_F \leq \rho_i$, for each data point $\mathbf{x}_i$, ρ_i was calculated by the Frobenius norm of the difference between the replicates generation covariance matrix and the sample covariance matrix. Then model **(SVM-SOCP-Cov)** was used in this case.

For the performance measures, NomErr and OptErr as discussed in Sect. 4 are used here. Now each test data point $\mathbf{x}_i$ has 50 replicates $\mathbf{x}_{i_k}$. The class label y_i^{pr} is decided by the majority label of the replicates $\text{sign}(\mathbf{w}^\top \mathbf{x}_{i_k} + b)$. For the OptErr, the sample mean and sample covariance are used to calculate ε_{opt}. We randomly partitioned 20 % of the data as training and the remaining 80 % as the test. The results over 20 runs are shown in Table 3 and Fig. 9.

The results show that since OptErr considers the probability of misclassification even when the predicted label is correct, OptErr is always bigger than NomErr. NomErr and OptErr do not have common trends among the four cases (one not considering the estimation error, and three considering different estimation errors). For $\boldsymbol{\mu}_i \in [\boldsymbol{\mu}_i^-, \boldsymbol{\mu}_i^+]$, the running time is greater than the other cases, this makes sense since the model **(SVM-SOCP-Mu1)** is more complicated than the other models and it also has the largest robust region.

6 Conclusion

This paper studied SVM when uncertainties exist in data. Chance constraint is to ensure the small probability of misclassification of the uncertain data. Robust optimization is to guarantee an optimal performance when the worst-case scenario constraints are still satisfied. This paper obtained equivalent SDP and SOCP reformulations for the robust chance-constrained SVM when the second-order moment information of the uncertain data are known. Optimization problems with such kind of data uncertainties can be reformulated and proved similarly. The SDP reformulation also provides the potential for further extension to joint chance constraints where the data points not only have their own distributions, but also correlated with each other. Numerical experiments showed the equivalence while the SOCP model works more efficiently. The geometric interpretation of the SOCP model shows that the model actually transformed the chance constraints into robust ellipsoid regions. The estimation errors are discussed and geometrically interpreted when the mean vector and covariance matrix are estimated from the data. The models considering estimation errors are also proposed for different cases.

For further research, the numerical algorithms on big data is a potential direction and application. Currently, because of availability of such types of data sets, we can only perform limited numerical experiments on some existed data with our own modifications. In the future, through proposed approaches, more meaningful results could be obtained on different data sets.

Acknowledgments We are grateful to Danial Kuhn and Berç Rustem for their valuable discussions. We would like to thank the anonymous reviewers for their helpful comments. Research was conducted at National Research University, Higher School of Economics, and supported by RSF grant 14-41-00039.

References

Abe, S. (2010). *Support vector machines for pattern classification*. Berlin: Springer.

Ben-Hur, A., & Weston, J. (2010). A users guide to support vector machines. In O. Carugo & F. Eisenhaber (Eds.), *Data mining techniques for the life sciences* (pp. 223–239). Berlin: Springer.

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

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

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

Bi, J., & Zhang, T. (2005). Support vector classification with input data uncertainty. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17: Proceedings of the 2004 conference*. Cambridge: MIT Press.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

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

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

Fan, N., Sadeghi, E., & Pardalos, P. M. (2014). Robust support vector machines with polyhedral uncertainty of the input data. In P. M. Pardalos, M. G. C. Resende, C. Vogiatzis, & J. L. Walteros (Eds.), *Learning and intelligent optimization* (pp. 291–305). Berlin: Springer.

Ghaoui, L. E., Lanckriet, G. R., & Natsoulis, G. (2003). *Robust classification with interval data*. Technical report UCB/CSD-03-1279, Computer Science Division, University of California, Berkeley.

Ghaoui, L. E., Oks, M., & Oustry, F. (2003). Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, *51*(4), 543–556.

Isii, K. (1960). The extrema of probability determined by generalized moments (i) bounded random variables. *Annals of the Institute of Statistical Mathematics*, *12*(2), 119–134.

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.

Marshall, A. W., & Olkin, I. (1960). Multivariate chebyshev inequalities. *The Annals of Mathematical Statistics*, *31*(4), 1001–1014.

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* (pp. 369–374). World Scientific and Engineering Academy and Society (WSEAS).

Pólik, I., & Terlaky, T. (2007). A survey of the s-lemma. *SIAM Review*, *49*(3), 371–418.

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.

Tian, Y., Shi, Y., & Liu, X. (2012). Recent advances on support vector machines research. *Technological and Economic Development of Economy*, *18*(1), 5–33.

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*(7), 757–781.

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

Trafalis, T. B., & Gilbert, R. C. (2007). Robust support vector machines for classification and computational issues. *Optimization Methods and Software*, *22*(1), 187–198.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999.

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

Xanthopoulos, P., Guarracino, M. R., & Pardalos, P. M. (2014). Robust generalized eigenvalue classifier with ellipsoidal uncertainty. *Annals of Operations Research*, *216*(1), 327–342.

Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2012). *Robust data mining*. Berlin: Springer.

Yakubovich, V. A. (1971). S-procedure in nonlinear control theory. *Vestnik Leningrad University*, *1*, 62–77.

Zymler, S., Kuhn, D., & Rustem, B. (2013). Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, *137*(1–2), 167–198.